Spectroscopic techniques applied to the real time monitoring and control of the production processes of the pharmaceutical and food industry

Doctoral Dissertation
Doctoral Program in Materials Science and Technology (31st Cycle)

Spectroscopic techniques applied to the real time monitoring and control of the production processes of the pharmaceutical and food industry

Gentian Gavoci
* * * * * *

Supervisors
Prof. F. Geobaldo
Prof. F. Savorani

Doctoral Examination Committee:
Prof. G. Zeppa, Referee, Dipartimento di Scienze Agrarie, Forestali e Alimentari (Italy)
Prof. M. Arlorio, Referee, Università del Piemonte Orientale (Italy)

Politecnico di Torino
April 2019

I hereby declare that, the contents and organisation of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Gentian Gavoci
Turin, April 2019

Preface and Objectives

This dissertation is about the use of spectroscopic techniques, mainly near infrared spectroscopy (NIRS), in combination with multivariate data analysis for the real time monitoring and control of the production processes of the pharmaceutical and food industry. The expected final product quality is the main task to accomplish for the industrial organizations taking also into account the costs and environmental impact. Thus far, the quality assessment of manufactured products is performed primarily on a post-production based testing, using an off-line laboratory strategy. This approach may result in products that do not meet the requested quality characteristics and should be disposed or reworked. The implementation of the strategy based on the Process Analytical Technology (PAT) concept, accepted as an effective tool for process monitoring and control, represents an innovative solution to avoid unwanted consequences stemming from the abovementioned quality control approach. The PAT paradigm involves an active process control, starting from the quality control of incoming raw materials and through the continuous process control, leading to semi-finished or final products within specifications having little variation with respect to the critical quality parameters. Spectroscopic sensors, in combination with computational analysis (multivariate data analysis), are regarded as the most advantageous process analysers for PAT successful implementation, moving the quality measurements closer to the process via at-line, on-line and in-line strategies. PAT, allowing real time collection, analysis and sharing of production process data, becomes a powerful tool in the manufacturing step included in the management of the value chain over the life cycle of pharmaceutical and food products as addressed by Industry 4.0.

This thesis consists of three mainly parts. The first part contains four chapters.

The first chapter is an introduction related to Industry 4.0 emphasising, the advantages for manufacturing process due to a set of constituents belonging to design principles and technology trends, and the benefits the pharma- and food-industry can obtain.

Chapter 2 describes near infrared spectroscopy (NIRS) as the spectroscopic technique predominantly employed in this research activity to collect spectral data of solid and liquid samples according to diffuse reflectance, transmittance and transflectance modalities.

Chapter 3 presents the multivariate data analysis as essential computational instruments to get qualitative and quantitative responses from the spectral data

acquired by a secondary analytical technique like NIRS. The principles of designs of experiments (DoE) are also briefly discussed.

Chapter 4 includes the principles and components of the PAT approach as well as the strategies of NIRS analysis. At the beginning of the chapter is also presented an introduction of the most recent approach which takes advantage of the PAT benefits, known as Quality by Design (QbD).

The second part (Chapter 5) contains three industrial research case studies that I have faced during my PhD period.

The first case study (Section 5.1), related to a pharmaceutical industry, has been focused on two main objectives: 1. The development of chemometric models, for identity confirmation and identification (classification) of incoming raw materials, which allow to compare the NIR spectrum of a material (known or unknown) with the spectra of known materials employed previously for qualitative modelling; 2. The development of a regression model based on the PLS method in order to predict the quantity of the active ingredient DHA (docosahexaenoic acid) in a semi-finished or end solid product.

The second case study (Section 5.2), involved a firm of vegetable oils where the research activity was focused on four aims: 1. The assessment of the shelf life of vegetable oils; 2. Quality evaluation of extra virgin olive oils as a function of the storage time; 3. The effect of two different presses on five vegetable oils produced starting from their seeds; 4. The evaluation of cold-pressed linseed oil oxidative stability when subjected to accelerated oxidation.

The last case study (Section 5.3) concerns the coffee industry and has, as its main aim, the classification of green coffee beans according to their geographical origin using NIRS and chemometrics.

The third part (Chapter 6) contains the conclusions and the future perspectives of this research approach.

Acknowledgment

I would like to thank you Prof. Francesco Geobaldo e Prof. Francesco Savorani who gave me the opportunity to carry out my PhD experience at the Department of Applied Science and Technology (DISAT) of Politecnico di Torino.

I am also very grateful to Dr. Alessandro Giraudo and Dr. Nicola Cavallini for their generous support during my research work.

Contents

1. Introduction 1

1.1 Industry 4.0 1

1.1.1 Design principles 3
1.1.2 Technology trends 5

1.2 Pharma and food industry 4.0 9
1.3 References 10

2. Near infrared spectroscopy 13

2.1 Fundamentals 13
2.2 Instrumentation 16

2.2.1 Process instruments 18
2.2.2 Miniaturized instruments 19
2.2.3 NIR imaging instruments 20

2.3 Acquisition techniques 22

2.3.1 Diffuse reflectance 22
2.3.2 Transmittance 22
2.3.3 Transflectance 23

2.4 Applications in food and pharmaceutical areas 24
2.5 HSI in agro-food and pharmaceutical sector 25
2.6 References 26

3. Multivariate data analysis (Chemometrics) 28

3.1 Design of experiments (DoE) 28

3.1.1 Full factorial designs 30
3.1.2 Fractional factorial design 32
3.1.3 Plackett Burman designs 33

3.1.4 Mixture designs 33
3.1.5 Designs for multivariate calibration 33
3.2 Exploratory analysis 34
3.2.1 Pre-processing techniques 34
3.2.1.1 Classical pre-processing methods 34
3.2.1.2 Signal correction methods 35
3.2.2 Principal component analysis (PCA) 37
3.2.3 Outlier detection 39
3.3 Classification techniques 41
3.3.1 Soft Independent Modelling of Class Analogy (SIMCA) 42
3.3.2 Partial Least-Square Discriminant Analysis (PLS-DA) 42
3.4 Calibration techniques 43
3.4.1 Multiple linear regression (MLR) 44
3.4.2 Principal component regression (PCR) 44
3.4.3 Partial least squares (PLS) 44
3.5 Validation techniques 45
3.5.1 Cross-validation 45
3.5.2 Test set validation 46
3.5.3 Bootstrap 46
3.6 Model evaluation criteria 46
3.7 Process Monitoring and Control 47
3.7.1 Multivariate Statistical Process Control (MSPC) 47
3.8 References 48
4. Process Analytical Technology and Quality by Design 50
4.1 Quality by Design (QbD) 50
4.2 Process Analytical Technology (PAT) 52
4.2.1 PAT principles 53
4.2.2 PAT components 54
4.2.2.1 Critical product quality attributes and process parameters 54
4.2.2.2 Process dynamics and sampling 55
4.2.2.3 Spectroscopic techniques and imaging 55

4.2.2.4. Chemometrics 56
4.3 NIR analysis strategies 57
4.3.1 Off-line testing 57
4.3.2 At-line testing 58
4.3.3 On-line testing 58
4.3.4 In-line testing 58
4.3.5 Non-invasive testing 59
4.4 PAT implementation in food and pharmaceutical industries 59
4.5 References 61
5. Case studies research 64
5.1 Pharmaceutical industry 64
5.1.1 Introduction 64
5.1.2 Materials and methods 65
5.1.2.1 Raw materials and semi-finished product 65
5.1.2.2 Data collection 70
5.1.2.3 Multivariate data analysis 70
5.1.3 Results and discussion 71
5.1.4 References 81
5.2 Vegetable oil company 82
5.2.1 Introduction 82
5.2.2 Materials and methods 83
5.2.2.1 Materials and designs 83
5.2.2.2 Data collection 85
5.2.2.3 Multivariate data analysis 86
5.2.3 Results and discussion 87
5.2.4 References 94
5.3 Coffee industry 95
5.3.1 Introduction 95
5.3.2 Materials and methods 96
5.3.2.1 Coffee samples 96
5.3.2.2 Data collection 96
5.3.2.3 Multivariate data analysis 97

5.3.3 Results and discussion....98
5.3.4 References....100
6. Conclusions and future perspectives....102
6.1 Conclusions....102
6.1.1 Pharmaceutical industry....102
6.1.2 Vegetable oil company....103
6.1.3 Coffee industry....103
6.2 Future perspectives....104
6.2.1 Pharmaceutical industry....104
6.2.2 Vegetable oil company....104
6.2.3 Coffee industry....105

List of Tables

Table 3-1. Experimental runs to be performed considering three factors each one at two levels (modified from: Brereton, 2007). 31

Table 3-2. Design matrix related to the experiment presented in Table 3-1 (modified from: Brereton, 2007). 32

Table 5-1. List of the botanicals and other solid raw materials with the highest number of lots purchased by the company (*Raw materials subjected to exploratory analysis and classification purposes that have been included in this work). 66

Table 5-2. Incomplete list of the solid raw materials employed to make the product. 69

Table 5-3. Working conditions of the mechanical expeller pressing and hydraulic press. 84

Table 5-4. Results of continent-based Partial Least Squares-Discriminant Analysis (PLS-DA) on NIR spectra collected by DISAT and UNIMI after the application of different mathematical pre-processing: EFF obtained in calibration (CAL), cross-validation (CV) and prediction of the external test set (TS); Raw: raw spectral data without any pre-processing except MC; SNV: standard normal variate; MSC: multiplicative scatter correction; d1: first derivative; d2: second derivative. 99

List of Figures

Figure 1-1. Schematic representation of technology trends and design principles involved in the passage towards Industry 4.0 (modified from: Ghobakhloo, 2018). 3

Figure 2-1. Main chemical bonds responsible for the NIR spectra and location of their vibrational frequencies (modified from: Osborne, 2006). 15

Figure 2-2. Schematic depiction of the several parts making up the NIR spectrometer (modified from: Blanco, 2002) 17

Figure 2-3. Absorption, reflection and transmission of NIR radiation following the interaction with the sample. 22

Figure 2-4. Diffuse reflectance and transmittance phenomena involving radiation. 23

Figure 3-1. Geometric depiction of the experimental design illustrated in Table 3-1 (modified from: Brereton, 2007). 31

Figure 3-2. Raw spectral data (a) and mean centring (b) pre-processing outcome starting from the raw data of ”Vitamin A”. 35

Figure 3-3. Raw spectral data (a), MSC (b), SNV (c) and Savitzky-Golay second derivative (d) pre-processing outcomes starting from the raw spectra of ”Xangold”. 36

Figure 3-4. Visualization of the PCA decomposition structure 38

Figure 3-5. Loadings plot (a) and scores plot (b) of the raw material xangold, as obtained from the first and second principal components. 39

Figure 3-6. Outlier (ginkgo biloba dry extract 6%) detection by using scores plot and 95% of confidence limit. 40

Figure 3-7. Influence plot of spectral data representing two Ginkgo biloba dry extracts according to a confidence limit of 95%. 41

Figure 3-8. PLS-DA model of two classes depicted by the $\boldsymbol{Y}$ column vector (modified from: Brereton, 2014). 43

Figure 3-9. Graphical visualization of the PLS regression for a three component model (modified from: Wold, 2001). 45
Figure 4-1. Schematic representation of the monitoring and control process involving a manufacturing unit operation (modified from: van den Berg, 2013)..54
Figure 4-2. Representation of the several strategies which can be used to test the product throughout the process. 57
Figure 5-1. Buchi FT-NIR instrument used to collect the spectral data from the abovementioned samples. 70
Figure 5-2. PCA scores plot depicting the limits of the known group of libramed samples based on the Mahalanobis distance. 72
Figure 5-3. PCA scores plot of arginine base and hydrochloride using SNV and MC as pre-processing techniques. 73
Figure 5-4. PCA scores plot of orange and lemon aroma using SNV and MC as pre-processing techniques. 73
Figure 5-5. PCA scores plot of anhydrous and monohydrate citric acid using SNV and MC as pre-processing techniques. 74
Figure 5-6. PCA scores plot of alpha and beta galactosidase using Savizky-Golay (window size: 19; polynomial order: 2; 2nd derivative) and MC as pre-processing techniques. 74
Figure 5-7. PCA scores plot of potassium citrate and sorbate using MSC and MC as pre-processing techniques. 75
Figure 5-8. PCA scores plot of two grades mannitol, according to particles size, using SNV and MC as pre-processing techniques. 75
Figure 5-9. PCA scores plot of thioctic acid in the two forms (Matris fast and retard) using SNV and MC as pre-processing techniques. 76
Figure 5-10. PCA scores plot of ubidecarenone in the two forms using SNV and MC as pre-processing techniques. 76
Figure 5-11. PCA scores plot of xangold in the two forms using EMSC and MC as pre-processing techniques. 77
Figure 5-12. PCA scores plot of the three forms of ginkgo biloba dry extracts using SNV and MC as pre-processing techniques. 77
Figure 5-13. PCA scores plot of the three forms of melissa dry extracts using SNV and MC as pre-processing techniques. 78
Figure 5-14. PCA scores plot of soybean dry extracts, purchased from two suppliers, using SG (window size: 19; polynomial order: 2; 2nd derivative) and MC as pre-processing techniques. 78

Figure 5-15. Figure 5-15. PCA scores plot of four different dry extracts using SG (window size: 19; polynomial order: 2; 2nd derivative) and MC as pre-processing techniques. 79
Figure 5-16. Ginkgo biloba 24% pre-processed spectra (a) and their subsequent prediction on the previously developed PLS-DA model (b). 79
Figure 5-17. PLS linear regression model depicting the predicted versus the actual concentration of DHA in the various mixtures ranging from 5% to 10% according to the DoE explained in Section 5.1.2. 80
Figure 5-18. Prediction of six NIR spectra representing two final products manufactured in two different times. 81
Figure 5-19. Expeller press equipped with a thermocouple and an in-house assembled portable FT-NIR instrument equipped with a fibre optic transflectance probe. 85
Figure 5-20. FT-NIR MPA spectrometer from Bruker optics (a) and Varian Cary 5000 UV-Vis Spectrophotometer (b). 86
Figure 5-21. NIR spectrum of olive oil with the assignment of the main absorption bands. 86
Figure 5-22. Visible impact of hemp oil arising from NEON and LED light exposure (a) and the spectra of the three oils collected in the visible region (b). 87
Figure 5-23. PCA scores plot of NIR spectral data of the three vegetable oils. 88
Figure 5-24. PCA scores plot of visible spectral data of the three vegetable oils over the six months of storage under controlled conditions. 88
Figure 5-25. Scores plot of the collected spectral data of extra virgin olive oils with a different storage time (gap of six months of storage). 89
Figure 5-26. NIR spectral data resulting from extra virgin olive oil stored in tank S704 over the five months. 90
Figure 5-27. PCA scores plot of extra virgin olive oil during the five months of storage. 90
Figure 5-28. Evolution of the oil temperature, detected on the top and bottom, within tank S704. 91
Figure 5-29. PCA models of hempseed, sunflower and pumpkin oils yielded starting from their seeds. 91
Figure 5-30. PCA models of walnut and linseed oils yielded by the two presses. 92
Figure 5-31. Overall PCA scores plot of the mean spectra of each replicate (three replicates for hemp, sunflower, linseed and pumpkin oils; two replicates for walnut oil). 92

Figure 5-32. NIR spectra of linseed oil acquired by using the bench-top instrument, right after pressing under two different speeds, and after accelerated oxidation. 93
Figure 5-33. PCA scores plot of linseed oil samples before and after accelerated oxidation in stove. 94
Figure 5-34. PCA scores plot of samples having different country of origin (a) and the average of SNV American and Asian NIR spectra of green coffee beans (b). 99

Chapter 1

Introduction

1.1 Industry 4.0

Industry 4.0 is a term originated in Germany in 2011 which depicts the world's fourth industrial revolution following the three previous industrial revolutions which extended across for nearly 2 centuries. The first industrial revolution, that began at the end of the seventeenth century, developed mostly in England due to the inventions in the textile and iron metallurgy sectors and the availability of raw materials: coal, used for the operation of steam engines and iron, employed in the siderurgic and mechanical sectors with emphasis in the cast iron production. Cotton imported from the colonies led to the growth of the cotton industry. The second industrial revolution spanned mostly from 1870's to 1970's and was characterized by several scientific and technological discoveries, inventions and innovations as well as the use of new sources of energy that enabled radical transformations in industry. Henry Ford reorganized the entire factory around the assembly line officialising mass production and reducing production time and costs. The third industrial revolution was driven by the use of Information and Communication Technology (ICT), electronics and robotics to produce more automation in industrial manufacturing. Small scale and smart computer tools become mature enough to deliver services and IT infrastructure through smart networks. It is also being introduced the concept of lean production which is a systematic method aimed to reduce the waste, costs, stocks and to provide different kinds of products. Compared to the previous revolution the employees are characterized by a higher education and flexibility. The industrial sector expanded and developed all around the world along with the induced issues like climate change, pollution and sustainability.

The definition created into the framework of the "Industry Platform 4.0", started by companies, industrial associations, the Federal Ministry for Economic Affairs and Energy and the Federal Ministry of Education and Research in Germany in 2015 reports that: "The term Industrial 4.0 stands for the fourth industrial revolution, a new stage in the organization and management of the entire value

chain over the life cycle of products. This cycle addresses the increasingly individualized customer requirements and extends from the idea of the development and manufacturing, the delivery of a product to the customer up to the recycling, including associated services. The basis is the availability of all relevant information in real time through the networking of all entities involved in the value creation as well as the ability to derive, from the data, the optimal value flow at any time. By linking people, objects and systems, dynamic, real-time and self-organizing, cross-company value-added networks emerge, which can be optimized according to different criteria such as cost, availability and resource consumption" (Wilkesmann, 2018). Industry 4.0 can be regarded as the implementation of cypher physical systems (CPS) with particular emphasis on the customers who are actively involved within the industrial production systems (Wan, 2015). The four essential constituents of Industry 4.0 are indentified with 1) the internet of things (IoT); 2) the CPS; 3) the internet of services (IoS) and 4) the intelligent and self-organizing factory (Hermann, 2016).

According to the European Parliamentary Research Center (2015) the importance of industrial sector for the European Union (EU) economy decreased by one third over the last four decades. The reduction of value added by manufacturing is due to the growing of industry in countries with lower manpower costs and digital manufacturing is expected to stimulate economic growth increasing productivity and added value. However, there are great uncertainties about the social effects of this new phenomenon (European Parliament website).

Many scholars are convinced that Industry 4.0 is an imminent event and the impact is compared to the impact the Internet had as an important technology. In order to remain competitive in the chaotic market, manufacturers must pave the way for digitized manufacturing changing in this way manufacturing processes, business models and outcomes (Ghobakhloo, 2018). The successful transition towards Industry 4.0 is reached by producing and implementing a strategic roadmap including strategic and technological steps in the direction of a full digital organization (Vogel-Heuser, 2016; Sarvari, 2018). Technology roadmapping is widely employed by modern businesses as a structure to sustain the research and development of coming technologies maintaining a possible competitive advantage (Lee, 2013).

The successful transition from traditional to digital manufacturing involves a profound understanding of the characteristics of Industry 4.0 as a precondition for the progress of the strategic and technological roadmap. According to many scholars, design principles and technology trends have been considered the fundamental constituents of Industry 4.0. The study made by Ghobakhloo (2018) reviews these fundamental constituents of Industry 4.0 analysing their advantages for manufacturing processes identifying a set of key design principles and technology trends related to Industry 4.0 (Ghobakhloo, 2018). Design principles address the problem relevant to the identification and implementation of Industry 4.0 scenarios by offering an arrangement of knowledge and describing the constituents of this phenomenon (Hermann, 2016) sustaining, in this way, professionals in developing suitable solutions. The rise of the new digital industrial

technology is permitted by technology trends that pertain to the advanced digital technological innovations (Gilchrist, 2016). The following figure (Figure 1-1) depicts the architecture of Industry 4.0 taking into account these technology advancements and design principles.

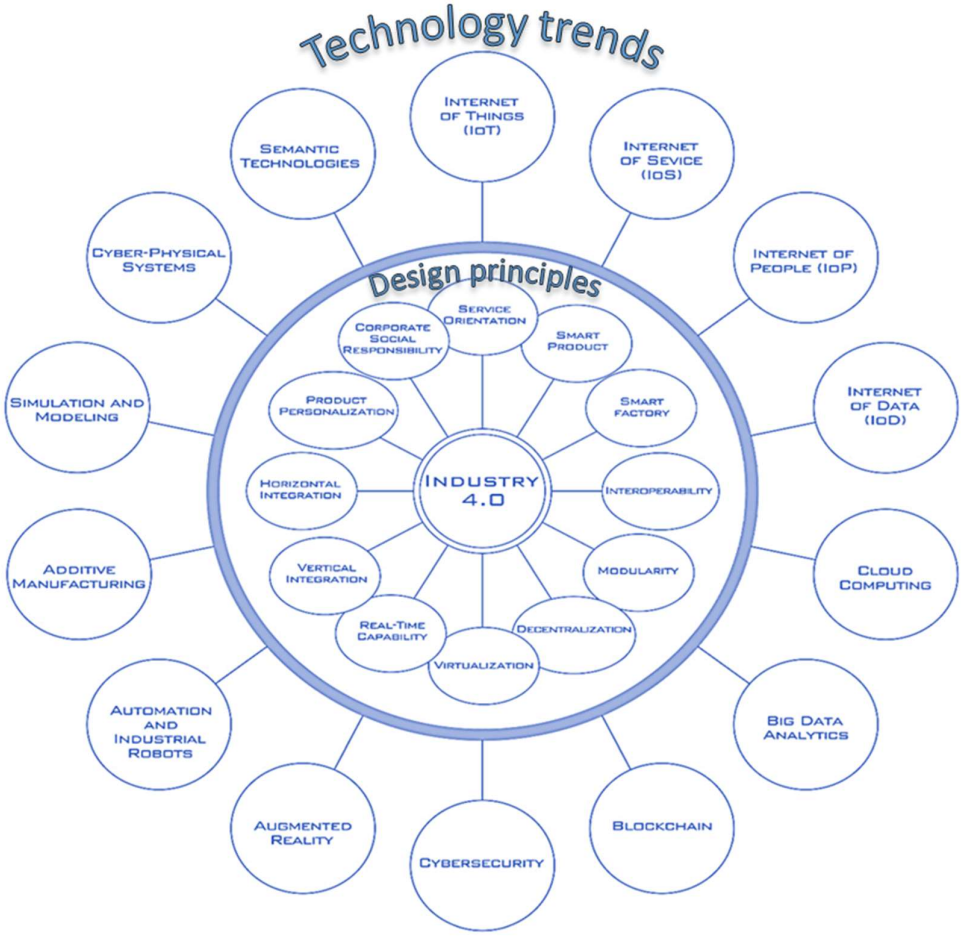

Figure 1-1. Schematic representation of technology trends and design principles involved in the passage towards Industry 4.0 (modified from: Ghobakhloo, 2018).

A brief description of the abovementioned design principles and technology trends has been carried out as follows.

1.1.1 Design principles

Service orientation. This principle is mainly related to Manufacturing as a Service (MaaS) and Product as a Service (PaaS) business models. MaaS refers to the diffusion of IoT and cloud manufacturing along with interconnectivity among manufacturers which enable Organizations to readily share their production demand and capability in order to use a common connected manufacturing structure to yield goods. This means collaboration among Organizations to perform complicated

tasks (Ghobakhloo, 2018). The capability of monitoring and managing in real time and in a remote way fabricated products by IoS allows to build business models in which a service is provided instead of a product and the user would benefit of a shared resource and pay this service on the basis of its actual use. An appropriate connection among people and smart objects via IoS enables to manufacture products that satisfy customer's specifications.

Smart product. Smart products within the Industry 4.0 framework are products equipped with embedded sensors able to self-process, memorize and transmit data, provide information about their identity and current status as well as communicate inside the industrial environment. During the manufacturing phase the various production steps that lead to the end product as well as the maintenance activities are communicated because of the product capability in computing algorithms and machine learning (Schmidt, 2015).

Smart factory. Smart factory depicts a smart, highly digitized and productive manufacturing organization with connected machines, devices and systems which results in the prediction of equipment downtime and minimized waste. IIoT, IoP and WoT technologies in the smart factory allow not only the communication among smart objects but also between them and human resources in order to make real time decisions (Ghobakhloo, 2018).

Interoperability. The interoperability is related to the ability to accurately and quickly communicate, operate, manage the data and share of information via IIoT, IoS, IoP, and WoT among all constituents of the manufacturing industry such as relevant technologies, smart products, human resources, etc. and with the manufacturing partners like suppliers, customers and other interested parties (Gilchrist, 2016). This results in a better decision making.

Modularity. The modularity concerns the transformation from rigid systems, linear manufacturing and planning and inflexible business models towards a ductile environment able to manage the ever-changing requests in the supply chain, market situations and other relevant requirements that need flexibility (Gilchrist, 2016).

Product personalization. Product personalization is related to the product development based on changing customer needs and preferences, enabled by developments in technology trends and identified by evaluating customers' behaviour. Moreover, the manufacturers should predict customers' behaviour and not just fulfil their current needs by foreseeing market evolution (Ghobakhloo, 2018). This new manufacturing paradigm follows mass production (relatively inexpensive and limited variation of products) and mass customization where the product variation is augmented significantly (Berry, 2013).

Decentralization. The decentralization refers to the autonomous operation and decision making of several constituents of the smart factory taking into account the common organization goal (Gilchrist, 2016). However, human decisions are still very important and so only some decisions can be fully automated (www.i-scoop.eu).

Virtualization. This design principle is related to the concept of "digital twin" that consists in the ability to create virtual models, starting from sensor data collected from the entire value chain of the real organization (warehouse, systems,

processes and products), that provide information from the physical objects. The assessment of the process yielding the product, as well as the customer satisfaction, is accomplished by the digital twin of the smart product as it allows to get a full virtual footprint along the product life-cycle. Real time capability has an essential role in virtualization because it includes data acquisition, analysis of data and decision making in real time and also the detection in real time of cyber security threats (Ghobakhloo, 2018).

System integration. System integration incorporates either the vertical networking of layer upon layer of technologies and manufacturing systems (factories, products, etc.) or the horizontal integration related to the overall connection through the value chain with the aim of distributing the wanted functionality and creating value through networks institution (Ghobakhloo, 2018).

Corporate social responsibility. It is part of the enterprise's business model and is mainly related to the manufacture labour and environment rules. Some believe that Industry 4.0 is going to reduce the availability of jobs whereas others think that technology will create more jobs compared to those that will be cancelled. The organization, in order to be part of the industrial revolution, must train their employees with the aim to develop skills useful for Industry 4.0. Sustainable design of products and processes, sustainable manufacturing, increased performance of employees and implementation of green business models are achievable within the Industry 4.0 context improving environmental sustainability (Ghobakhloo, 2018).

1.1.2 Technology trends

Internet of things. The IoT refers to physical objects connected to the Internet and equipped with electronics, sensors, software and actuators able to exchange data with other connected objects (e.g. sensors able to catch process data and then to send the information to people and other products). The connected objects will enable the real time optimization of production processes and economic activities considerably reducing pollution and resource consumption. In the Industry 4.0 framework the IoT is usually mentioned as Industrial Internet of Things (IIoT) which indicates the industrial application of IoT (Wang, 2016) as different actors have addressed this paradigm on their finalities, interests and backgrounds (Atzori, 2010). IIoT pertains either the connected physical objects or the digital depiction of processes, products and manufacturing facility (Jeschke, 2017). Smart and connected tools related to IIoT capture and communicate data in a more accurate and consistent way compared to humans, whereas the interaction between humans and objects enabled by IIoT results in the remote monitoring and control of devices (Almada-Lobo, 2016). Moreover, the knowledge management (KM) IoT concerns the accomplishment of a wider network among people, objects and systems (Roblek, 2016). The continuous gathered data by monitoring machines and systems through KM IoT are combined with smart algorithms, allowing companies to get access to new information useful in decision making processes and creating innovations (Wilkesmann, 2018). The greater combination of data among

companies, suppliers and clients can reconsider the demand for intermediary parties (Porter, 2014).

This technology can also play an important role in the preventative maintenance of equipment subjected to wear as well as in the assessment of functionality and usage of products and permitting an improved planning capacity due to the inventory monitoring (Bughin, 2015). The IoT allows individual identification of products which will be connected to details about their provenance, use and destination with no need to correlate products and information flows. For the common users there are also some risks related to the IoT, such as the improper use of data and privacy protection.

Internet of services. IoS is related to the systematic use of the Internet to create and provide a high number of services based on the Product as a Service (PaaS) business model (Ghobakhloo, 2018). Previously separated available services on the Internet will be combined to create a more extended network of services with high added value with the involvement of several actors that can take advantage reciprocally. IoT represents the basic structure to support IoS, ensuring the connection between sevices whereas the possibility to store and analyse large amounts of data arising from the offered services allows to increase the value of the service itself. The company selling a service, thanks to the IoS, can make a real time connection to the IT system of the customer to assess and monitor the condition of the provided service. In this way the producer can remotely intervene in a prompt manner to anticipate possible incoming issues. Moreover, there is also the possibility to directly collect data from the customer to assess the performance of the product, either to deliver proactive and preventive maintenance (Leminen, 2012) or sending the information to the research and development unit to yield a product with higher performances.

Internet of data (IoD). IoD refers to the analysis of data on a larger scale and its aim is to record all the activities and monitor the data entities along their life cycles. The huge amount of data related to the high number of objects in the IoT domain should be effectively transferred, stored, managed and processed by the IoD. Accordingly, the IoD is useful as a constituent of IoT, IoS and the internet of People (IoP) and it is considered comparable with database management systems (Ghobakhloo, 2018).

Cloud computing. Cloud computing development is due to the continuous recent progress in virtual technology, hardware, distributed computing and service delivery beyond the Internet (Oliveira, 2014). It is a technology offered as a service that enables to exploit software and hardware resources through remote server. It can be used to create new services, new applications, to store huge amounts of data and analyse the data to obtain strategic models and production plans and to perform continuous dat backups and promptly recover needed information. Cloud computing is fast and versatile enabling the enterprises to get the sought information in a short time and practically everywhere. The implementation of cloud manufacturing allows the integration of allocated manufacturing assets and the institution of an interactive infrastructure across organization spots allocated in different areas (He, 2015).

Big data. Big data refer to architectures, technologies and analytics that allow Organizations to extract value and hidden knowledge through generation, collection, storage and analysis of huge volumes of complex and various data. Organizations can use big data analytics to predict the outcome of an operation and the actions needed to get the optimal results (LaValle, 2011). Predictive models can also be built, starting from successful visualization (holograms, augmented and virtual reality methods), analysis and sharing of data originating from product development and manufacturing processes, and implemented in various smart manufacturing plants (Demartini, 2016). Pushing the boundaries of knowledge by using big data analytics, the Organizations can augment both the accuracy of knowledge and the accuracy of decisions and actions to be taken. The transformation of a product to a service (servitization), predictive maintenance, increase of product customization, agile production processes and more effectively management of supply chain are achieved through big data analytics. Big data analytics can also be used by Organizations to assess potential suppliers according to their previous performance and to establish performance indicators for managing the suppliers (Ghobakhloo, 2018).

Blockchain. Blockchain is a communication protocol which identifies a technology based on the distributed ledger that enables the creation and management of a large distributed database (e.g. for the management of transactions). Blockchain as an open and transparent technology, able to guarantee immutability and incorruptibility, decentralised, shared, encrypted with strict security rules allows transparent, secure, reliable and quick private or public solutions (Underwood, 2016). Within the paradigm of Industry 4.0, blockchain can work as a distributed ledger among several constituents of smart factories, suppliers and customers to develop reliable and autonomous relation (Ghobakhloo, 2018).

Augmented reality (AR). AR transforms huge quantities of data in images, overlapping the developed images with the real world. The proper visualization of the progress of activities enabled by AR ensures the improvement of production processes, quality control, maintenance and assistance. AR technology has also been implemented by modern manufacturers in support of personnel training, quality management and product design optimising the design phases, forestalling the issues related to the product, reducing the product development costs (Ghobakhloo, 2018).

Automation and industrial robotics. This combination depicts two interacting technologies as the under way trend towards automation in manufacturing requires a growing need for industrial robots. Automation and industrial robotics permit improvements in the efficiency of processes by optimizing the operation, higher quality and reliability, reduced costs and time, increased reliability, reduced waste and better space usage (Ghobakhloo, 2018).

Cybersecurity. This building block represents a critical technology in the implementation of Industry 4.0 since the cyber risks can be of different nature. Inside Industry 4.0 environment, everything is always connected through the Internet that means more susceptibility toward the external environment. Accordingly, the risk of stealing data, information and core know-how from the

companies increases significantly. The challenges of Industry 4.0 with regard to cybersecurity are related either to the ordinary information security or to the own privacy and security problems (Ghobakhloo, 2018).

Additive manufacturing (AM). AM describes the techniques that build objects by adding point to point and layer upon layer of material in accordance with the original model. The model, usually, generated by a CAD (computer aided design) software is transformed into a real object by using 3D printers starting from various materials of solid or liquid appearance (Strange, 2017). Industry 4.0 is going to reduce the distance between customers and suppliers allowing customers to address production orders to the manufacturing company in real time. The advantages related to the additive manufacturing such as the speed of production, freedom, accuracy, supply chain cost reduction and small-scale manufacturing tests point out that additive manufacturing can sustain the smart factory idea (Ghobakhloo, 2018).

Simulation and modelling. These technologies are already used for products, production processes and materials but in the future these techniques will be employed in plant operations in a larger scale, reflecting the physical world in a virtual model which can include machines, products and humans. Simulation and modelling allow manufacturers not only to avoid errors at an early stage avoiding costs for plant operators, but they can also be employed to optimise a manufacturing plant during in progress daily operations. According to industrial reports, the virtual testing of entire production systems is a common goal for many manufacturers (Ghobakhloo, 2018).

Cyber physical systems (CPS). CPS is regarded as a technology with the potential of creating value along the three dimensions of digitalized manufacturing (i.e. smart product, smart manufacturing and business model). It is a collection of different technologies (sensors, actuators and decentralized intelligence) able to generate an autonomous, intercommunicating and intelligent system able to facilitate integration among different and distant subjects. This system enables data generation and collection, computation and aggregation of previously acquired data and supporting decision making. The physical feature of products, systems and processes connected to the network among them and integrated with elements (embedded sensors, actuators) provided with computation, storage (memory) and communication capability is joined to their virtual or digital depiction. Smart factory, made feasible through CPS implementation, ensures efficiency related to resource utilization as all the manufacturing steps are integrated and are manufactured solely the products required by the customer due to the ability to monitor the markets (Report smart manufacturing (sCorPiuS project)). According to this project (research project funded by the European Union) there are six classes of benefits arising from CPS implementation.

- "New data driven services and business models" refer in particular to the managerial area with new opportunities of business allowing the company to be closer to the customers need.
- "Data based improved products" is related to the advantages arising from the product digitization. The product is able to communicate inside and outside the firm sharing information, enabling a better understanding of

processes and services, augmenting the added value for the final client. The information obtained from the organization about the product utilization consents to get the customers feedback in real time creating services and making products tailored for each client.

- "Closed loop manufacturing" includes the benefits related not only to the company but to the whole value network including suppliers and customers by integrating their own data and feedback with the aim to create zero waste supply chains.
- "Cyberized plant/Plug & Produce" considers the benefits that CPS brings at the shop floor that makes easy the optimization and management of operations, validates flexible and reconfigurable production system scenarios, plant self-recovery, self-learning, self-analysis and the product traceability during manufacturing.
- "Next step production efficiency" refers to the achievement of a more efficient production, able to make the manufacturing of small lots sustainable, speeding up and enhancing the production processes precision.
- "Digital ergonomics" includes the advantages arising from the introduction of tools and cyber physical technologies that consider a faster knowledge transfer process, a work experience improvement and a reduction of operational complexity.

Semantic technologies. These technologies facilitate integration, interoperability and analysis of data, processes and services providing reference models as well as information- and knowledge-sharing among various components of Industry 4.0. Semantic technologies have an important role concerning the management of things, tools and services whereas the several components and their constituents follow a unified and consistent model. In this way the process of integration of new components turn out to be more rapid and so the communication in a networked manufacturing (Ghobakhloo, 2018).

1.2 Pharma and food industry 4.0

Regarding the pharmaceutical production processes, the upcoming and existing technologies related to Industry 4.0 paradigm enable the transition from batch-related production, where the process cannot be successfully controlled, to sustainable continuous production (unbroken flow of raw materials and final products) with on(in)-line quality monitoring and control. The integration of Process Analytical Technologies (PAT) within Industry 4.0 enables the collection, analysis and sharing of real time quality data across the supply chain by using sensors in continuous manufacturing. Sensors can also collect information in real time about the environment surrounding the production line. Whether during the continuous production processes the quality standards of any product constituents are out of specifications or not, it will lead to real time decisions affecting the subsequent steps. The Quality by Design (QbD) approach, introduced in the

following chapter, helps the companies to implement continuous manufacturing ensuring successful quality control by designing the whole production process and fosters them to continuously improve the quality management. A more effective process monitoring, from raw materials to product utilization, arises from the improvement in communication due to Industry 4.0 technologies which enables the share of information flow in real time. The implementation of technologies that compose Industry 4.0, the increased knowledge and the improved control of processes in the manufacturing sector of the pharmaceutical industry, will support quality improvement and the prediction of production processes. The vertical and horizontal integration of Industry 4.0 can promote mass customization through smart production lines and the different actors of the entire supply chain are more aware of personalised customer demands (Ding, 2018). Technologies such as 3D printing can enable the continues production of rapid and personalized pharmaceutical products (Goole, 2016).

The components of Industry 4.0 have been applied in food industry in the optic of mass customization related to yoghurt production. The relevant selection of raw materials, the development of a model that allows continues production of several products on a non-reconfigurable production line, the virtualization of products, packaging, labelling and the preventive maintenance of the production line have been made possible by QR code technology, supervisory control and data acquisition (SCADA) system, radio frequency identification (RFID) for internal product tracking and other elements (Simon, 2018).

The digitalization of the food industry is a new subject, with limited scientific literature, gaining interest only over the last years. Demartini (2018) proposed a paper based on interviews and literature review that emphasizes the development of a useful structure and a model for the digitalization of manufacturing processes in food sector by using the Manufacturing Value Modelling Methodology (MVMM). According to this work, the selected enabling technologies for food manufacturing digitization consist of cyber physical production systems (CPPS), IIoT, cloud and additive manufacturing, big data analytics and holograms. The suggested model, that should be validated considering more than one case study, is mainly qualitative and does not permit sufficient quantitative evaluations.

1.3 References

Wilkesmann, M., & Wilkesmann, U. (2018). Industry 4.0–organizing routines or innovations? *VINE Journal of Information and Knowledge Management Systems*, *48*(2), 238-254.

Wan, J., Cai, H., & Zhou, K. (2015, January). Industrie 4.0: enabling technologies. In *Proceedings of 2015 international conference on intelligent computing and internet of things* (pp. 135-140). IEEE.

Hermann, M., Pentek, T., & Otto, B. (2016, January). Design principles for industrie 4.0 scenarios. In *2016 49th Hawaii international conference on system sciences (HICSS)* (pp. 3928-3937). IEEE.

www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI%282015%29568337. Industry 4.0: Digitization for productivity and growth.

Ghobakhloo, M. (2018). The future of manufacturing industry: A strategic roadmap toward Industry 4.0. *Journal of Manufacturing Technology Management*, *29*(6), 910-936.

Vogel-Heuser, B., & Hess, D. (2016). Guest editorial Industry 4.0–prerequisites and visions. *IEEE Transactions on Automation Science and Engineering*, *13*(2), 411-413.

Sarvari, P. A., Ustundag, A., Cevikcan, E., Kaya, I., & Cebi, S. (2018). Technology roadmap for Industry 4.0. In *Industry 4.0: Managing The Digital Transformation* (pp. 95-103). Springer, Cham.

Lee, J. H., Phaal, R., & Lee, S. H. (2013). An integrated service-device-technology roadmap for smart city development. *Technological Forecasting and Social Change*, *80*(2), 286-306.

Gilchrist, A. (2016). *Industry 4.0: the industrial internet of things*. Apress.

Wang, S., Wan, J., Zhang, D., Li, D., & Zhang, C. (2016). Towards smart factory for industry 4.0: a self-organized multi-agent system with big data based feedback and coordination. *Computer Networks*, *101*, 158-168.

Atzori, L., Iera, A., & Morabito, G. (2010). The internet of things: A survey. *Computer networks*, *54*(15), 2787-2805.

Jeschke, S., Brecher, C., Meisen, T., Özdemir, D., & Eschert, T. (2017). Industrial internet of things and cyber manufacturing systems. In *Industrial Internet of Things* (pp. 3-19). Springer, Cham.

Almada-Lobo, F. (2016). The Industry 4.0 revolution and the future of manufacturing execution systems (MES). *Journal of innovation management*, *3*(4), 16-21.

Roblek, V., MešKo, M. and Krapež, A. (2016), "A complex view of industry 4.0", Sage Open, Vol. 6 No. 2, pp. 1-11.

Porter, M. E., & Heppelmann, J. E. (2014). How smart, connected products are transforming competition. *Harvard business review*, *92*(11), 64-88.

Bughin, J., Lund, S., & Manyika, J. (2015). Harnessing the power of shifting global flows. *McKinsey Quarterly*, *7*(1), 1-13.

Leminen, S., Westerlund, M., Rajahonka, M., & Siuruainen, R. (2012). Towards IOT ecosystems and business models. In *Internet of things, smart spaces, and next generation networking* (pp. 15-26). Springer, Berlin, Heidelberg.

Miranda, J., Mäkitalo, N., Garcia-Alonso, J., Berrocal, J., Mikkonen, T., Canal, C., & Murillo, J. M. (2015). From the Internet of Things to the Internet of People. *IEEE Internet Computing*, *19*(2), 40-47.

Oliveira, T., Thomas, M., & Espadanal, M. (2014). Assessing the determinants of cloud computing adoption: An analysis of the manufacturing and services sectors. *Information & Management*, *51*(5), 497-510.

He, W. and Xu, L. (2015), "A state-of-the-art survey of cloud manufacturing", International Journal of Computer Integrated Manufacturing, Vol. 28 No. 3, pp. 239-250.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT sloan management review*, *52*(2), 21.

Demartini, M., Pinna, C., Tonelli, F., Terzi, S., Sansone, C., & Testa, C. (2018). Food industry digitalization: from challenges and trends to opportunities and solutions. *IFAC-PapersOnLine*, *51*(11), 1371-1378.

Underwood, S. (2016). Blockchain beyond bitcoin. *Communications of the ACM*, *59*(11), 15-17.

Strange, R., & Zucchella, A. (2017). Industry 4.0, global value chains and international business. *Multinational Business Review*, *25*(3), 174-184.

https://scorpius-project.eu/ - Documents & publications – Scorpius Papers and articles.

Schmidt, R., Möhring, M., Härting, R. C., Reichstein, C., Neumaier, P., & Jozinović, P. (2015, June). Industry 4.0-potentials for creating smart products: empirical research results. In *International Conference on Business Information Systems* (pp. 16-27). Springer, Cham.

Berry, C., Wang, H., & Hu, S. J. (2013). Product architecting for personalization. *Journal of Manufacturing Systems*, *32*(3), 404-411.

Ding, B. (2018). Pharma industry 4.0: Literature review and research opportunities in sustainable pharmaceutical supply chains. *Process Safety and Environmental Protection*, *119*, 115-130.

Goole, J., & Amighi, K. (2016). 3D printing in pharmaceutics: A new tool for designing customized drug delivery systems. *International journal of pharmaceutics*, *499*(1-2), 376-394.

Simon, J., Trojanova, M., Zbihlej, J., & Sarosi, J. (2018). Mass customization model in food industry using industry 4.0 standard with fuzzy-based multi-criteria decision making methodology. *Advances in Mechanical Engineering*, *10*(3), 1687814018766776.

Chapter 2

Near infrared spectroscopy

2.1 Fundamentals

Since the discovery of near infrared (NIR) electromagnetic radiation by William Herschel several decades have passed before encountering the first studies in analytical implementations. NIR spectra characteristics, the absence of mathematical tools as well as the slow advances in techniques that make use of NIR radiation brought to this gap. NIR spectroscopy is based on the exploitation of electromagnetic radiation at wavenumbers usually between 12800 and 4000 cm^{-1} corresponding to the wavelength range between 780 and 2500 nm where the absorption bands are mostly due to overtones and combinations of fundamental vibrations (stretching and bending) (Blanco, 2002).

According to Bokobza (1998), the absorption of infrared radiation by molecules causes the vibration of individual bonds which can be described through the ideal harmonic oscillator model. For the diatomic molecule the energy of equally spaced levels of harmonic oscillator based on the quantum mechanical treatment is given by:

$$\boldsymbol{E_{vib}} = \hbar\boldsymbol{\nu}\left(\mathbf{v} + \frac{\mathbf{1}}{\mathbf{2}}\right) \qquad \text{Equation 2.1}$$

where:
$\hbar$ = the Plank's constant
v = the vibrational quantum number
ν = the vibrational frequency expressed as:

$$\boldsymbol{\nu} = \frac{\mathbf{1}}{\mathbf{2}\boldsymbol{\pi}}\left(\frac{\boldsymbol{k}}{\boldsymbol{\mu}}\right)^{\mathbf{1/2}} \qquad \text{Equation 2.2}$$

where:
k = the classical force constant
μ = the reduced mass of the bonding atoms

Transitions from a vibrational energy level to another occur when the absorption of radiation causes a dipole moment change which implies that only the transitions involving heteronuclear diatomic molecules are possible. Moreover, in the harmonic oscillator only the transitions between consecutive energy levels are

allowed because the vibrational quantum number can only vary by one unit ($\Delta v = \pm 1$).

According to Boltzmann distribution the most likely transitions occurs between the ground vibrational level (v = 0) and the first vibrational level (v = 1) with higher energy since the ground level is more populated at room temperature. This transition called fundamental transition is characteristic of the middle infrared region and has the following energy (Blanco, 2002):

$$\Delta \boldsymbol{E_{vib}} = \Delta \boldsymbol{E_{rad}} = \hbar \boldsymbol{\nu} \qquad \text{Equation 2.3}$$

The bands related to the allowed transitions between consecutive excited vibrational levels (v = 1 → v = 2, etc.) are called "hot bands" which are weaker and have the same frequency as that of fundamental transition according to the harmonic oscillator.

The model based on the harmonic oscillator does not consider the Coulombic interaction or the dissociation of bonds involved in the molecule. Moreover, based on experimental observations the "hot bands" are characterized by frequencies that differ from the frequency of the fundamental band. The model that can better describe the behaviour of molecules is based on anharmonic oscillator where the vibrational energy levels are not equally spaced and the energy gap decreases with increasing the vibrational quantum number (v) according to the following expression (Blanco, 2002; Bokobza, 1998):

$$\boldsymbol{\Delta E_{vib}} = \hbar \boldsymbol{\nu}[\mathbf{1} - (\mathbf{2v} + \boldsymbol{\Delta} \mathbf{v} + \mathbf{1})\boldsymbol{y}] \qquad \text{Equation 2.4}$$

where:
y = the anharmonicity factor

The anharmonicity enables the transitions between non adjacent vibrational levels where $\Delta v > 1$. These multilevel vibrational energy transitions give rise to overtone bands that roughly take place at multiples of the fundamental vibrational frequency ($\Delta v = 2$→first overtone, $\Delta v = 3$→second overtone…). The probability of these transitions decreases when the vibrational quantum number increases and the overtone bands are much weaker in intensity than the fundamental bands since their transitions are less likely and occur between 12800 and 5000 cm^{-1}. NIR combination bands are generated when the absorbed radiation modifies at the same time the vibrational energy levels of two or more interatomic bonds. This results in vibrational interactions expressed as the sums of multiples of each interacting frequency which appear between around 5200 cm^{-1} and 4000 cm^{-1} (Blanco, 2002).

The main bands appearing in the NIR region are related to chemical bonds containing the hydrogen atom and other elements with low atomic weighs (Figure 2-1). This includes C-H, O-H, N-H and S-H bonds whereas the bands related to bonds such as C=O and C-C are much weaker or even missing compared to the previous ones. The presence of the hydrogen as the lightest atom in bonds results

in a greater deviation from harmonicity with higher intensity overtone bands (Osborne, 2006).

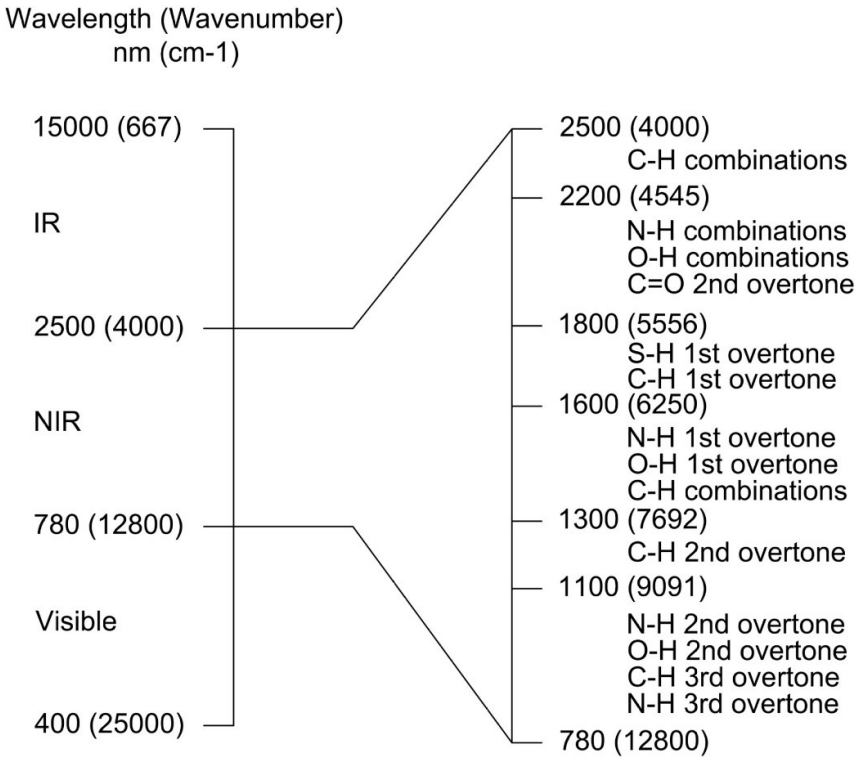

Figure 2-1. Main chemical bonds responsible for the NIR spectra and location of their vibrational frequencies (modified from: Osborne, 2006).

The spectra arising from NIR analysis consist of overlapping, broad and weak bands (10-100 times weaker than the corresponding bands arising from the absorption of the mid infrared radiation) with difficulties in the visible interpretation of spectra due to the lack of selectivity. Usually, a chemometric tool is employed to overcome this drawback in order to relate spectral information to sample properties. There are also benefits related to the low absorption coefficient because it allows the NIR radiation to enter deeper into the sample as well as the direct analysis of samples without further pre-treatment. In this manner it is not necessary to use solvents or other chemicals that have negative environmental impacts. Moreover, the NIR analysis are fast and time saving since the spectra are obtained in few seconds (Blanco, 2002).

The information contained in the NIR spectrum is of chemical (chemical composition of the sample) and physical origin such as particle size, crystal form, viscosity, etc. Elemental compositional changes at the molecular level, such as the substitution of light atoms with heavier atoms, will determine different NIR spectra even if the bonds related to the substituted atoms are not directly involved in the generation of bands. This is due to the effect of the new bonds on the an-harmonic vibrational modes and strength of the remaining bonds. Increasing the strength of

the chemical bond and/or decreasing the mass of the connected atoms, the relevant spectral absorption band(s) will shift towards higher wavenumbers. When analysing solid samples there is a linear relationship between the absorbance and the concentration of the absorbing species only when the range of concentration is limited as the scattering effect arising from the interaction of the radiation with solid samples affects the expected linear relation between absorbance and concentration. One way to reduce or remove the scattering effect consists in the mathematical treatment of the raw spectra (Pasquini, 2018).

2.2 Instrumentation

The need for fast analyses and instrument flexibility in order to suit for different types of samples brought to the development of NIR instrumentation that can merge several devices as shown in Figure 2-2. Depending on the wavelength selection NIR spectrometers can be discriminated in discrete wavelength and whole spectrum acquisition. The discrete wavelength instruments usually use filters that select relatively broad bands or light emitting diodes (LED) that generate narrow bands. These tools have been used in analysis with analytes absorbing at particular spectral ranges since only a few wavelengths interact with the samples. Acousto-optic tunable filters (AOTFs) are another type of wavelength selection dispersive monochromators that perform more reliable, rapid and reproducible wavelength scans than the grating instruments. NIR instruments based on whole spectrum usually incorporate dispersive grating monochromators and more recent systems based on Fourier Transform. Due to their advantages in flexibility compared to the discrete ones, they can be used in a wider range of applications (Blanco, 2002).

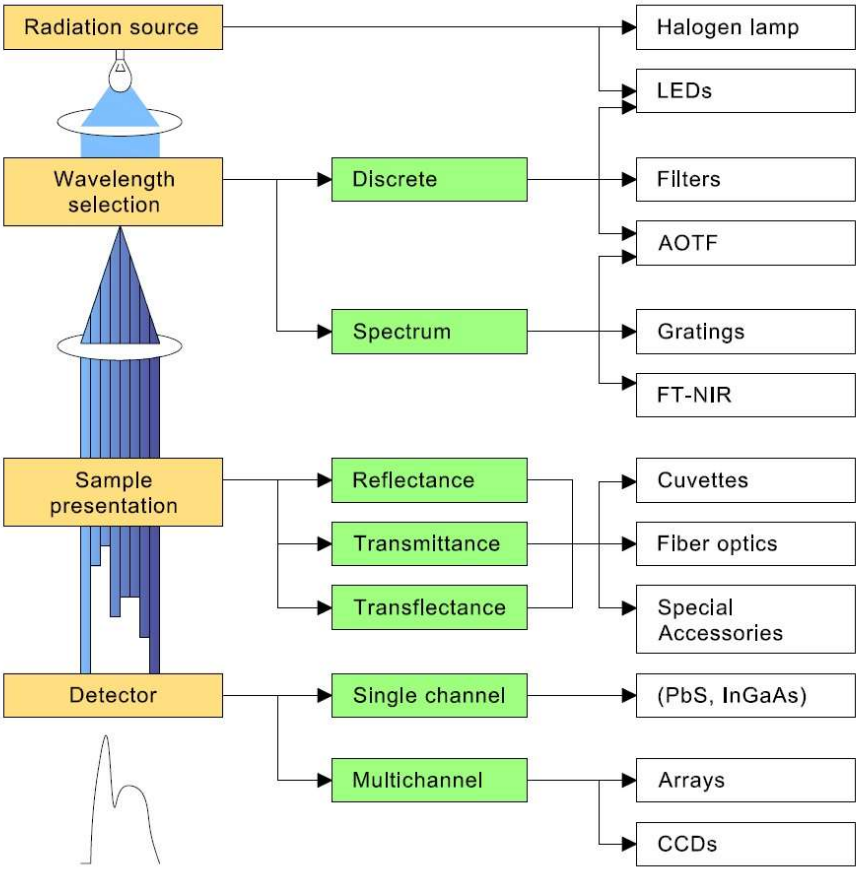

Figure 2-2. Schematic depiction of the several parts making up the NIR spectrometer (modified from: Blanco, 2002).

There has been an evolution concerning the laboratory bench NIR instrumentation from the equipment employing filters and those based on diffraction gratings to the Fourier transform spectrometers equipped with interferometer (FT-NIR). The interferometer represents the most important component of the spectrometer, which consists of a beamsplitter and two mirrors (one of which is movable), with the function of modulating the light. FT-NIR spectrophotometers depict the most widespread commercial technology as laboratory bench NIR tools besides their robustness, durability and the progressive reduced size. Within a controlled laboratory ambient the above-mentioned technology con offer very high spectral resolution (seldom necessary for spectra arising from solid samples but useful when analysing other mixtures having almost identical chemical features where little changes in the spectra are important to build qualitative and quantitative multivariate regression models), very good signal-to-noise ratio (SNR), quick spectra collection and spectral range covering the whole NIR region. The quality of NIR spectra provided by these instruments turn out to be very good in all the known acquisition modes, that is diffusive reflectance, transmittance and transflectance (Pasquini, 2018).

Detectors incorporated in the NIR technology can be divided in two classes: single channel and multichannel detectors. The single channel detectors are mainly based on semiconductors such as PbS and InGaAs. Detectors based on PbS have been the first ones employed in the NIR instrumentation and today are still used even though the detectivity shown is lower and the response time longer compared to InGaAs semiconductors. This last single channel detector is the most used in the modern FT-NIR spectrophotometers achieving good performance over the entire NIR spectral range even if it loses a little in SNR. Both the FT-NIR and dispersive based instruments display very good SNR, often surpassing 10,000:1. Detectors based on mercury cadmium telluride are also employed, even though they exhibit a lower performance over the entire spectral range. The multichannel detectors include diode arrays, in which elements are arranged in rows, and charged coupled devices (CCDs) where elements are arranged in planes. These systems allow to record the spectral information more rapidly by collecting many wavelengths at once. NIR imaging spectroscopy arises from this kind of detectors where the spectra collected by cameras provide a three dimensional image by which the shape and size of the sample is recorded besides the determination of the spatial composition and the wavelength employed (Blanco, 2002; Pasquini, 2018).

NIR instruments equipped with fiber-optic probes provide fast spectra acquisition by selecting the appropriate mode depending on the sample characteristics that is, usually, diffusive reflectance for solids, transmittance for liquids and transflectance for opaque liquids and emulsions. Fiber-optic probes are usually an integrated part of process instruments.

2.2.1 Process instruments

Due to its advantages such as the non-invasive testing and the speed of analytical response (a few seconds), NIR process instruments have been employed as analytical techniques in several manufacturing sectors as integrating part of the PAT framework over the last years. PAT concept, which will be further described in detail in Chapter 4 of this thesis, has been mostly implemented in pharmaceutical industries where the production process is monitored and controlled by NIRS taking into account the instrument specifications. The operation of process instruments must be autonomous and the accomplishment of all the necessary procedures to guarantee the quality of spectral information must be automatic. The instrument performance and the sampling process tests must be carried out throughout the device operation to assure reliable results (Pasquini, 2018).

An important aspect related to process instruments has to do with the collection speed of the spectral data. The sampling time, which can differ from a few seconds to some minutes, must be around ten times faster compared to the response time of the process. FT instruments are mostly used for on-line process monitoring after attentive installation of the spectrophotometer and are located at a certain distance from the production process line. Moreover, since the interferometer represents the most susceptible part of the instrument, improvements have been made to mitigate

the negative vibration effects on it. These instruments are equipped with optical fibres which are able to convey the radiation from the instrument to the process stream and then to send it back to the spectrometer after its interaction with the sample. Sometimes, when needed and if feasible, the sample is withdrawn from the line and transferred to the NIR instrument. Both the FT and dispersive instruments are frequently used in process monitoring even though their limits related to the moving components. The best solution for in-line or on-line process monitoring and control concerns the employment of solid state NIR devices where the combination of planar or concave gratings with an array of sensors made of PbS, Si, or InGaAs is the best known way to mitigate moving parts, with the drawback concerning the lower spectral resolution and the shorter wavelength range. Spectrophotometers equipped with the abovementioned sensors prove to be robust and the resulting spectrum can be acquired in less than one second. Technologies like AOTF have advanced in order to build monochromators with no moving parts and also characterized by a high scanning speed which makes them suitable for applications in production processes (Pasquini, 2018).

2.2.2 Miniaturized instruments

Some miniaturized spectrophotometers recently developed are transforming the NIR spectroscopy technique because besides being light, not too expensive and of a small-size are also able to collect data in the field along the value chain including the acceptance of materials, production processes, delivery, use, etc. Several manufacturers provide instruments with different specifications where the more important characteristic concerns the spectral range (Pasquini, 2018).

Other small-sized spectrophotometers have a limited employment in the field due to their request for external radiation sources. Some portable NIR instruments can weight up to 1.5 kg and have halfway size. These spectrophotometers must include all the needed devices in order to work autonomously; most of them employ sensor arrays while others use MEMS (micro electro mechanical systems) as intermediate spectral component among other devices. A few manufacturers of portable instruments developed and produced miniaturized devices, employed to put together entire instruments, exploiting the progress in MEMS and microelectronics. In these instruments the radiation provenance used for diffusive reflectance measurements is internal and technologies like Bluetooth used for data transmission after collection are provided (Pasquini, 2018).

Viavi solutions company proposed a microNIR instrument which includes tungsten filaments as source of radiation, a filter for the selection of wavelengths and an array of InGaAs sensors which are sensitive to a limited interval of wavelengths. The signal to noise ratio of the collected spectra can be improved by increasing the number of scans that can be averaged. This kind of instrument has an optical resolution included between 15 and 20 nm with good results in different applications like the identification of incoming raw materials in pharmaceutical industries, the adulteration of biodiesel, etc (Pasquini, 2018).

The smallest NIR instrument produced up to now is based on the MEMS technology and make use of a Fabry Perot filter where an applied voltage selects the restricted radiation beam to be transmitted. Another type of FT-NIR instrument commercialized but not yet assessed for applications uses the MEMS technology to build a micro-interferometer (Pasquini, 2018).

A comparison made over the last years between the portable and bench instruments, including all areas of applications, revealed good performances of portable instruments although inferior comparing with the conventional ones. In order to standardize and transfer the models to various units, the stability and reproducibility assessment of miniaturised instruments presently in commerce is needed. The miniaturised NIR spectrometers compared to the bench instruments usually do not guarantee their robustness for a lengthy period of time because of the absence of an internal self-testing protocol. As the transfer of the model is important for NIRS applications at several points of analysis in production, delivery and usage, studies have proved that model transfer from bench to portable spectrophotometers is feasible (Pasquini, 2018).

When analysing non-homogeneous samples by miniaturized instruments a common drawback is the insufficient representativeness of the measure owing to the very limited sample surface area entering in contact with the probe, overall bringing to non-accurate outcomes. In such occurrences before developing any qualitative or quantitative chemometric model a sampling procedure must be planned and/or adequate sampling devices should be employed. Avoiding sample transfer by performing in-situ measurements can improve the accuracy. Comparing the analytical data collected by portable instruments with the data arising from reference methods could not be the right way to assess the utility of a developed method. Taking into account two aspects like accuracy and speed of analysis, the methods related to miniaturized instruments can be characterized by lower accuracy but they gain importance in the real time measurements compared to methods arising from classical spectrophotometers (Pasquini, 2018).

2.2.3 NIR imaging instruments

The devices employed as imaging instruments are usually spectral cameras able to get an image of the sample surface where a NIR spectrum is acquired and associated to each component of the spatial resolution (pixel) as opposed to the portable and bench spectrophotometers that provide a spectrum arising from the average composition of the explored sample area. The hypercube (3D cube) arising from the data collection and processing of the acquired spatial and spectral information consists of many slices arranged horizontally which coincide with the number of wavelengths and contain images related to the single wavelengths that make up the NIR radiation. The data set is described by the term “hyperspectral imaging” (HSI) when a large number of wavelengths is employed. Each pixel can be represented as a single sample with sizes which may vary from 1 x 1 to 600 x 600 μm and the NIR spectra acquired in each of these small samples create the HSI

data set (Pasquini, 2018). When only a few wavelengths are available throughout the whole spectrum it is possible to refer to "multispectral imaging" (MSI) technology.

Cameras based on the "focal planar array hyperspectral" employ technologies like the AOTF or MEMS integrated in variable Fabry-Perot interference filters as wavelength selection systems and an array of detectors of planar disposition. These instruments aimed to take a NIR hyperspectral image can employ different wavelength ranges of the NIR region. The main benefits related to this sort of camera include the achievement of qualitative and quantitative analytical tasks by choosing a limited range of wavelengths and the short time in getting the HSI. Moreover, the image must be acquired while the sample is steady (Pasquini, 2018).

Recently, a snapshot camera composed by a system which allow to choose the wanted wavelengths and a flat detector array has been suggested. These cameras are made of Fabry-Perot filters as tools for wavelength selection placed on a detector made of metal oxide semiconductor. This device can acquire up to 170 full images per second even if a limited range of NIR radiation has been evaluated. A possible employment of these instruments concerns the process control by imaging following the appropriate choice of the required wavelengths given the limited amount of spectral channels (Pasquini, 2018).

The advances in resolution and in spatial dimension are the main characteristics provided by HSI. Another benefit is the higher sensitivity of NIRS provided by HSI where low analyte concentrations are identified through imaging over the surface of the sample. Compared to the classical measurements based on reflectance, this technology allows the detection of lower amounts of adulterants or contaminants due to its analytical sensitivity and selectivity. By employing HSI, wide sample areas are quickly explored improving sample representativeness by providing a mean spectrum which arise from wider areas compared to conventional NIRS. The HSI technology embedded in cameras has also the potential for in-field applications in order to monitor larger areas with interest in the agricultural and environmental sectors (Pasquini, 2018).

The large amounts of data arising from HSI technologies can be processed by using the present-day micro-computers. Multivariate data analysis including pre-processing techniques can be used to get details about the sample composition, following the same way as classical NIRS, as the sample image can be considered like an ensemble of minute samples (pixels). The score values of the pixels forming an image on the PCA model can display the compositional distribution of the sample surface starting from the spectral data. Often, the term "chemical image" is referred to images deriving from samples where the scores of the pixels are used to rebuild it. The concentration of several analytes in each pixel area can be evaluated by combining multivariate curve resolution (MCR) with ordinary (OLS) or alternating least squares (ALS). The number and the concentration of compounds in each pixel can be estimated using the MCR-ALS by taking advantage of the analytes distribution, observed in the sample by imaging (Pasquini, 2018). All these data mining techniques will be further discussed in Chapter 3 of this thesis.

2.3 Acquisition techniques

According to the sample presentation there exist several ways to perform the collection of NIR spectra. The main strategies are delineated below according to Osborne, (2006).

2.3.1 Diffuse reflectance

When the electromagnetic radiation at the interface between two mediums is not involved in diffuse (scattering) phenomena the term specular reflectance is used. Diffuse reflectance (Figure 2-4), instead, is the process by which the incident unidirectional radiation is reflected in many directions. When the radiation is reflected in a diffuse way but without penetration into the sample the absorption process doesn't occurs. Instead, when the radiation penetrates the sample surface, it can be absorbed, reflected or transmitted (Figure 2-3). The modality of diffuse reflectance is usually used to acquire the NIR spectra of solid, powder and semiliquid samples. The absorbance (A), and so the concentration, is related to the diffusely reflected radiation (R) through a similar relation to that of the Lambert-Beer's law (A = log1/R). An important phenomenon which should be reduced or eliminated is the scattering of radiation as it affects the response values.

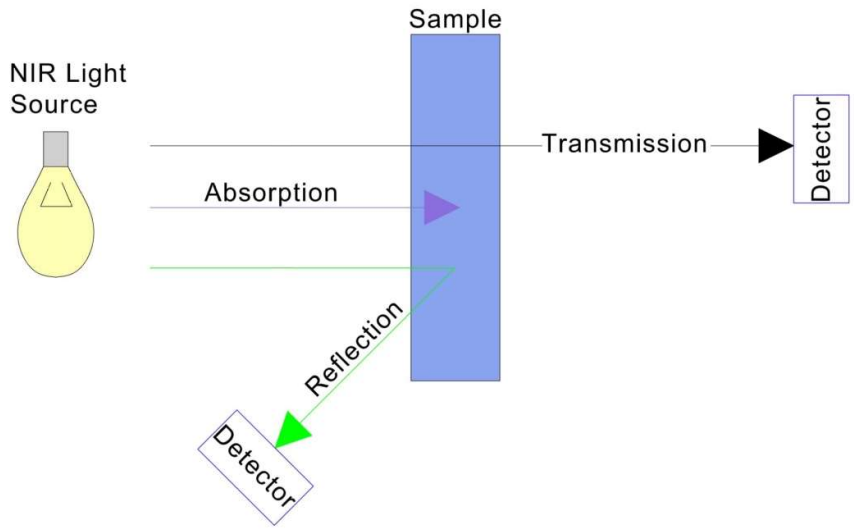

Figure 2-3. Absorption, reflection and transmission of NIR radiation following the interaction with the sample.

2.3.2 Transmittance

The transmittance mode is usually used to collect the spectra of transparent liquid materials (vegetable oils, vine, etc.) where the Lambert-Beer's law (A = εbc = log1/T), which put in relation the absorbance, the concentration (c) and the

transmitted radiation (T), is effectual as no scattering phenomena occur. For the acquisition of NIR spectra of vegetable oils there is no need to dilute the sample because of the low intensity absorption bands associated to overtones and combinations of fundamental vibrations.

In the presence of scattering the Lambert-Beer's law is not valid as the optical path (b) changes depending on the nature of the sample. Turbid liquids, semi-solids and solids being affected by light scattering can be analysed in diffuse transmittance (Figure 2-4). An example is that of the liquid whole milk in which the fat globules scatter the radiation changing the path length. Diffuse transmittance modality has been employed to collect the spectra of samples like grains, meat and cheese having a thickness of 1-2 cm. The appropriate arrangement of a device used for the whole grain analysis can also be employed for on-line measurements.

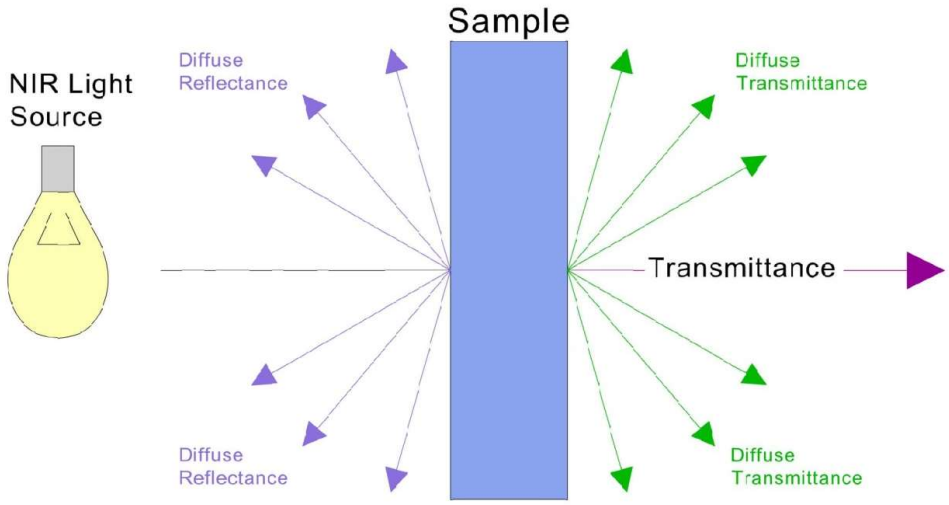

Figure 2-4. Diffuse reflectance and transmittance phenomena involving radiation.

2.3.3 Transflectance

Transflectance analysis arise from the combination of transmittance and reflectance. It is usually used with liquid samples where the radiation first is transmitted across the sample and then reflected by a suitable material (usually ceramic) placed beneath it and lastly transmitted back through the sample in order to reach the detector. The liquid sample is therefore crossed twice, and this allows a stronger signal to be acquired requiring smaller amounts of sample.

2.4 Applications in food and pharmaceutical areas

Based on the scientific literature, NIRS proved to be a useful technique for application in several fields due to its advantages compared to other analytical techniques.

Spectroscopic techniques including NIRS coupled with multivariate data analysis have been explored by several organizations for the quality assessment of extra-virgin and virgin olive oil. Major and minor constituents have been quantitatively evaluated along with the prediction of sensorial and technological characteristics. Another objective attained by using these technologies was the authentication of oils related to their production location (Gomez-Caravaca, 2016).

Coffee is one of the most consumed beverages and a high-quality product is necessary to be competitive by satisfying the clients. NIRS has the potential to provide information about the chemical components and the properties of coffee in a short time. This instrument can also be used for classification and authentication purposes as well as to assess sensory attributes. The knowledge about the evolution of the production process can be increased and for accomplishing this task further efforts must be made to transfer the knowledge in industrial applications (Barbin, 2014).

The detection of microbiological, chemical and physical hazards in various types of foods can be accomplished through NIR spectroscopy and imaging instruments combined with chemometrics. These techniques have also the potential to be applied for the food traceability and the discrimination of non-processed from the processed food. Besides the advantages over some other techniques, NIRS and imaging techniques are limited in the detection of trace amounts of chemicals which can be overcome by sample pre-treatment. The long-time collection of data is a drawback for spectral imaging instruments as it limits the on-line/real time applications (Fu, 2016).

Another potential field of application of NIRS concerns the adulteration of food and raw materials. In order to fingerprint incoming raw materials and ingredients as part of the process analytical technology (PAT) framework, it is required to move away from the targeted strategy (focused on the targeted prediction of the interested quality parameters). The high frequency monitoring of incoming raw materials by NIRS coupled with chemometrics should allow detecting deviations from the specifications increasing the speed of the production process (Sørensen, 2016).

NIRS has also been employed as a tool to assess the contamination of cereals by funghi and to estimate their secondary metabolites (mycotoxins). The outcomes arising from various studies showed the favorable employment of the technique for the identification of funghi and the assessment of specific levels of mycotoxins (Hossain, 2014) even though the inherent low sensitivity pose a strong limit.

One of the advantages of NIRS concerns the absence of sample preparation which is an important factor for real time on-line applications. Rapid analysis, however, is not the only requirement for successful on-line measurements as other factors such as temperature variation or moving samples affect the production process. Scientific studies have been made with the purpose of on-line

measurements but without clear results. Although numerous scientific studies claim the potential of NIRS for industrial on-line process monitoring and control, a limited number of scientific papers have been focused on the possibility of industrial on-line applications. Therefore, other studies should be performed to assess the NIRS application for real time measurements in the food industry taking into account the process conditions (Porep, 2015).

The pharmaceutical sector is another field of application of NIRS with studies aimed at the determination of the end point of mixing, granulation and drying manufacturing process. Qualitative and quantitative evaluations of pharmaceutical constituents are also performed by NIRS coupled with chemometrics. Final product physicochemical characteristics like porosity, hardness, compression, disintegration time and possible counterfeit have been monitored and controlled through spectroscopic determinations and the implementation of NIRS for the real time monitoring of pharmaceutical production processes will be a benefit for many industries (Jamrogiewicz, 2012).

The identification of raw materials and moisture content are extensively performed by NIRS which is considered as a standard method. Its potential applications in pharmaceutical companies span from the conformity check of incoming raw materials to the real time quality monitoring and control of the final pharmaceutical products. Along with other advantages, the employment of optical fibres makes it a prominent process analytical technology (PAT) technique enabling the optimization of the manufacturing chain (Luypaert, 2007).

The determination of the composition as well as the evaluation of natural products and the monitoring and control of production processes have been performed by NIRS. Its combination with chemometrics can enable to detect compositional characteristics which are even difficult to observe by conventional analytical techniques. Besides the advances in NIRS and chemometrics in order to fingerprint natural products, the endorsement of this technology in the abovementioned field is limited by the relevant training (Cozzolino, 2009).

2.5 HSI in agro-food and pharmaceutical sector

One of the main areas of application of HSI, concerns the agricultural sector, taking advantage of the spatial spectral potentialities. These instruments, for example, allow concomitant scanning of numerous kernels of various crops accelerating the analysis of each kernel. Information about several features like variety, possible contaminations as well as kernels classification can be obtained by integrating chemometrics. The detection of adulterants, such as melamine in milk powder by HSI can be obtained at lower concentrations compared with the conventional NIRS. Moreover, the HSI can give either surface or deep (up to a few millimeters) information of the interested sample. During the studies aimed to the adulterant detection it's worth considering the transformation process to yield the final product. Along this phase the properties (e.g. degree of crystallinity) of the interested analyte may change which result in different NIR spectra. Another

ambient of application of HSI is the pharmaceutical industry including works in the evaluation of homogeneity and isolation of active pharmaceutical ingredients in compressed products as well as other uses such as the distribution of polymorphic forms and investigations on the deterioration of tablets. HSI assisted by chemometrics is also employed in forensic studies for analysing biological fluids with identification purposes where non-destructive analyses are needed (Pasquini, 2018).

2.6 References

Blanco, M., & Villarroya, I. N. I. R. (2002). NIR spectroscopy: a rapid-response analytical tool. *TrAC Trends in Analytical Chemistry*, *21*(4), 240-250.

Bokobza, L. (1998). Near infrared spectroscopy. *Journal of Near Infrared Spectroscopy*, *6*(1), 3-17.

Osborne, B. G. (2006). Near-infrared spectroscopy in food analysis. *Encyclopedia of analytical chemistry: applications, theory and instrumentation*.

Pasquini C. (2018). Near infrared spectroscopy: a mature analytical technique with new perspectives – A review. Anal. Chim. Acta. 1026, 8–36.

A.M. Gomez-Caravaca, R.M. Maggio, L. Cerretani, Chemometric applications to assess quality and critical parameters of virgin and extra-virgin olive oil. A review, Anal. Chim. Acta 913 (2016) 1-21.

D.F. Barbin, A. Felicio, D.W. Sun, S.L. Nixdorf, E.Y. Hirooka, Application of infrared spectral techniques on quality and compositional attributes of coffee: an overview, Food Res. Int. 61 (2014) 23-32.

X.P. Fu, Y.B. Ying, Food safety evaluation based on near infrared spectroscopy and imaging: a review, Crit. Rev. Food Sci. Nutr. 56 (2016) 1913-1924.

K.M. Sorensen, B. Khakimov, S.B. Engelsen, The use of rapid spectroscopic screening methods to detect adulteration of food raw materials and ingredients, Current Opinion in Food Science 10 (2016) 45-51.

M.Z. Hossain, T. Goto, Near- and mid-infrared spectroscopy as efficient tools for detection of fungal and mycotoxin contamination in agricultural commodities, World Mycotoxin J. 7 (2014) 507-515.

J.U. Porep, D.R. Kammerer, R. Carle, On-line application of near infrared (NIR) spectroscopy in food production, Trends Food Sci. Technol. 46 (2015) 211-230.

M. Jamrogiewicz, Application of the near-infrared spectroscopy in the pharmaceutical technology, J. Pharmaceut. Biomed. Anal. 66 (2012) 1-10.

J. Luypaert, D.L. Massart, Y.V. Heyden, Near-infrared spectroscopy applications in pharmaceutical analysis, Talanta 72 (2007) 865-883.

D. Cozzolino, Near infrared spectroscopy in natural products analysis, Planta Med. 75 (2009) 746-756.

Chapter 3

Multivariate data analysis (Chemometrics)

Chemometrics is known as the branch of chemistry which studies the application of mathematical and statistical methods to the data arising from a system or from a chemical process (www.gruppochemiometria.it). The aim of chemometrics is:

- to design, select, and optimize experiments;
- to get as much as possible information from the system through data analysis;
- to transform the information in a graphical representation.

3.1 Design of experiments (DoE)

In order to understand and learn on systems and processes operation, the simple and attentive observation is important but not sufficient when it comes to understand what happens when the input factors are modified. Therefore, it is necessary to perform experiments in order to understand the system's output variations, i.e. the response variables, after changing the input variables, that is, to analyse the performance of systems and processes. According to Montgomery (2017), experimental design (or *design of experiments, DoE*) is used for various objectives which include:

- *Factor screening.* When studying a system or process there may be several factors that must be taken into account. It becomes important to understand which factors more affect the desired response values, that is, which factors are significant and should be studied in detail and which of them can be removed. The characterization of factors is carried out for new processes and systems and when the knowledge about the system is not sufficient to get the requested performance. A model can be developed based on the relation between the significant factors and the response and then used for decision making.
- *Optimization.* This step is usually performed after the identification of the significant factors with the aim of finding the levels of such factors which produce the desired outputs (e.g. maximize yield). As a screening experiment typically does not provide the optimum settings for the

significant factors the optimization experiment is a continuation of the factors' characterization.

- *Confirmation.* The confirmation experiment is usually carried out when the aim is to assess the coherence of the system's behaviour with existing theories or experience, like for the evaluation related to a new advantageous material (e.g. in terms of costs) which may be equivalent, in theory, to existing materials but should maintain the same characteristics that affect its use. The scale-up of a manufacturing process from the pilot plant to the full-scale production often requires a confirmation experiment.
- *Discovery.* The output arising from the exploration of new materials or new factors is usually determined by discovery experiments (e.g. the effect of new products in treating disease).
- *Robustness.* This kind of experiments are usually performed in order to understand the conditions that could bring to output responses with unacceptable variability. The objective is to set the levels of controllable factors into the system to minimize the response variability arising from factors that cannot be controlled very well.

A good DoE results in saving time and resources and leads to concrete benefits for experiments where mathematical models related to the system are built.

There exist many areas of applications that employ DoE methods. The implementation of these techniques in manufacturing process design and development as well as in process management may result in several benefits concerning process performance, development time and global costs. The design and development of new products (including product formulation) along with product improvement benefit from the application of experimental design as well. The advantages related to product realization may be the reduced time for the development of products, the augmented performance of the manufactured products and the inferior product costs (Montgomery, 2017).

Usually the relation between the experimental response and the factors is represented through a model. In order to get reliable predictions of the response, the number of experiments to be run is of a crucial importance. Statistical experimental design provides a guide about the number and the nature of experiments needed to be performed (Leardi, 2009). For a straight mathematical relationship between the response and a given factor, the degrees of freedom (D), which is a measure of how well the data fits the theoretical model, is given by the difference between the number of experiments (N) and the number of coefficients (P) of the model. In order to assess if the data are consistent with the theoretical model, the information increases with the higher number of the degrees of freedom (Brereton, 2007).

In order to get an idea about the experimental error it is necessary to repeat several times the experiment under the same conditions. It becomes very difficult to accomplish good predictions when the error is large. To calculate the degrees of freedom, the number of experimental replicates should also be taken into account. An experiment is usually good when the number of replicates is close to the number

of the degrees of freedom. The significance of factors and interactions between factors in models can be determined by the analysis of variance (ANOVA) available in the common statistical tools (Brereton, 2007).

The first step when implementing a design of experiment consists in creating an experimental set-up which depicts the experiments performed under different factors conditions. This can be done using a design matrix, that is a matrix in which each row denotes an experiment and each column represents one of the parameters arising from the model (regression model) that depends from the number of factors. The design matrix is used along with the measured response values to estimate the best fit coefficients of the regression model from which the significance of factors can be inferred (Brereton, 2007).

The choice of a probable design of experiment usually depends on some constraints (time, costs and equipment) and on the problem to deal with. In the following sections, a more thorough description of some of the most used designs has been done.

3.1.1 Full factorial designs

The study of the influence of two or more factors as well as their interactions on the response is usually carried out by factorial designs. Unlike the one-factor-at-a-time design where factors are varied one at the time regardless of interactions, in factorial designs all combinations of the levels of factors defined in the experiment are examined. When interactions are not considered, this can lead to erroneous conclusions. Furthermore, also the estimation of the effects of a factor at various levels of the other employed factors is obtained by using factorial designs. In order to perform a full factorial design, the number of experimental runs to make is $\mathbf{N = l^f}$ where f represents the number of factors, and l depicts the number of levels (Brereton, 2007).

An important class of general factorial designs is the two-level, k factors full factorial design ($\mathbf{2^k}$) where the levels can be either quantitative or qualitative. The 2^k designs are extensively used in factor screening experiments as it is possible to study the k factors by performing a low number of runs which results to be advantageous in the first steps of the research work. For instance, the 2^2 factorial design refers to a design with two factors where each factor is at two levels which means that the number of experimental runs to perform is four without replicates (Brereton, 2007).

When the purpose is to study the influence of three factors (A, B, C) and their interactions, each one taken at two levels, on the response values the number of experiments to run is eight ($2^3 = 8$). Denoting the low level of each factor (coded variable) as -1, the high level as +1 and the response y, the design set up can be depicted as in the Table 3-1.

Table 3-1. Experimental runs to be performed considering three factors each one at two levels (modified from: Brereton, 2007).

Experiment	A	B	C	Response
1	**-1**	**-1**	**-1**	**y1**
2	**1**	**-1**	**-1**	**y2**
3	**-1**	**1**	**-1**	**y3**
4	**1**	**1**	**-1**	**y4**
5	**-1**	**-1**	**1**	**y5**
6	**1**	**-1**	**1**	**y6**
7	**-1**	**1**	**1**	**y7**
8	**1**	**1**	**1**	**y8**

When replicates are performed then the total sum of the responses of each experimental replicate is considered in order to assess the factors' importance. The 2^3 can be visualized geometrically as a box (Figure 3-1) where each corner represents a given treatment combination.

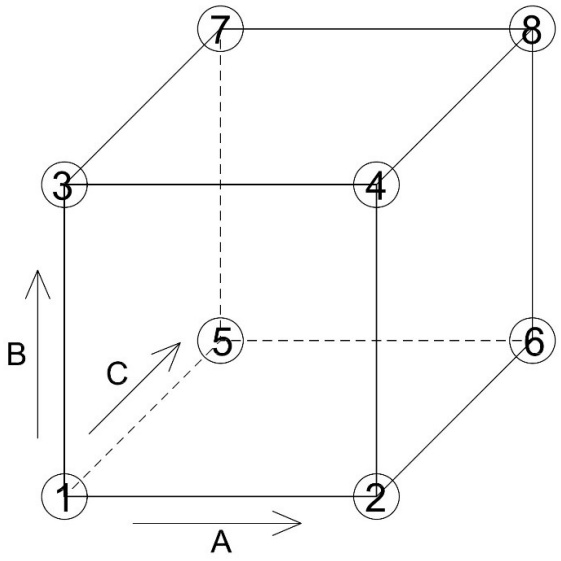

Figure 3-1. Geometric depiction of the experimental design illustrated in Table 3-1 (modified from: Brereton, 2007).

One possible way to understand the importance of factors and their interactions over the response is by estimating their corresponding parameters through the design matrix (Table 3-2). In order to optimize the system, the not relevant factors based on their parameters need to be eliminated.

Table 3-2. Design matrix related to the experiment presented in Table 3-1 (modified from: Brereton, 2007).

Intercept	A	B	C	AB	AC	BC	ABC
1	**-1**	**-1**	**-1**	**1**	**1**	**1**	**-1**
1	**1**	**-1**	**-1**	**-1**	**-1**	**1**	**1**
1	**-1**	**1**	**-1**	**-1**	**1**	**-1**	**1**
1	**1**	**1**	**-1**	**1**	**-1**	**-1**	**-1**
1	**-1**	**-1**	**1**	**1**	**-1**	**-1**	**1**
1	**1**	**-1**	**1**	**-1**	**1**	**-1**	**-1**
1	**-1**	**1**	**1**	**-1**	**-1**	**1**	**-1**
1	**1**	**1**	**1**	**1**	**1**	**1**	**1**
b_0	b_1	b_2	b_3	b_{12}	b_{13}	b_{23}	b_{123}

3.1.2 Fractional factorial design

A full factorial design involves an exponential augmentation of the number of experimental runs as the number of factors rises. In order to achieve the aim more time and resources are needed, which can become increasingly impracticable depending on the number of factors. In situations with a high number of factors the degrees of freedom associated with higher order interactions are likely to be insignificant or their physical meaning is difficult to understand. In this way only a fraction of the full factorial design can provide information about the main effects and the lower order interactions. These designs are called fractional factorial designs and are mainly used in screening experiments for product design and process design and improvement. The general formula of a two-level fractional design is $\mathbf{2^{k-p}}$. When p equals 1 the number of runs to be performed is reduced by half with respect to the complete factorial with consequently reduction of the amount of information. The number of experimental runs, for a three factor with each factor at two levels, equals four. The structure of the abovementioned design consists of the basic design which is a representation of the 2^2 factorial by adding the third factor that depicts the level interactions of the first and the second factors. This results in the *orthogonality* among the factors, that is the property that allow to differentiate the main effect of each factor varying distinctly (Montgomery, 2017).

3.1.3 Plackett Burman designs

To further reduce the number of experiments the Plackett-Burman designs can be employed. Through these designs the number of experimental runs to be carried out corresponds to the minimum number needed to study the factors of the system. These experiments are accomplished by using a "generator" in order to assure the *orthogonality* and exist for a selected number of factors (Brereton, 2007).

3.1.4 Mixture designs

Mixture designs require a set of factors, whose total percentage is constant and results to 100%, with the objective to find the optimal combination of components (factors) that yield the desired response. The several factors are interdependent and the mixture space, within which the experiments are depicted as points, turns out to be a triangle for three components, a tetrahedron for four components and so on (Brereton, 2007).

3.1.5 Designs for multivariate calibration

The conventional analytical methods (e.g. chromatographic techniques) are considered as time-consuming methods which makes them unsuitable when it is needed to get quick and continuous analytical responses. One way to obtain quick responses consists in collecting the spectra (e.g. NIR spectra) of samples that can be acquired in a few seconds and subjecting the spectra to computational methods in order to obtain calibrations that allow to extract quantitative information. A calibration model can be developed by collecting the spectra of samples with known concentrations of the compound of interest and, as a result, the future unknown concentration of the compound can be computed from the model. The data employed to build the calibration model is called training set. In order to provide reliable quantitative predicted responses of unknown samples, the design of the training set is very important. In the experimental design for the multivariate calibration the number of reasonable concentration levels is five (Brereton, 2007). The number of experimental runs to be performed for mixtures (made by more than one compound) equals $\mathbf{k} \times \mathbf{l}^{\wedge}\mathbf{p}$, where p is a whole number greater or equal to 2, k a whole number greater or equal to 1 and l corresponds to the number of concentration levels. Choosing five levels of concentrations with k and p at their lower levels, the number of experiments (mixtures) to perform according to this design is 25 (the resulting spectra). The benefit of this kind of design is that all the components of the mixture result orthogonal (Brereton, 2007).

3.2 Exploratory analysis

An exploratory analysis can provide information about the nature and the group membership of samples as well as information about the relationship between samples and variables. Usually, the multivariate data are subjected to a pre-processing step before exploratory investigation (O'Donnell, 2014).

3.2.1 Pre-processing techniques

Pre-processing methods are employed after the spectral data acquisition and before chemometric modelling. The spectra arising from NIR spectroscopic technique can be affected by non-linearities and baseline shifts which depend on the light scattering. The aim of pre-processing is to reduce or even remove these undesired variations (i.e. not related to the chemical nature of the sample) from the spectral data. For instance, Lorentz-Mie scattering is prevalent when the infrared radiation interacts with solid samples whose particle sizes are larger than the interested wavelength. The benefits arising from pre-processing can be of great importance during the steps of exploratory analysis, bi-linear calibration or classification modelling (Rinnan, 2009).

The abovementioned techniques can be distinguished between classical pre-processing methods and signal correction methods.

3.2.1.1 Classical pre-processing methods

These methods can be employed for a wide range of multivariate data including spectroscopic and process information data. Two of the most common methods are mean centring and scaling.

- *Mean centring:* The criterion consists in subtracting the variable mean of a specific position in the data matrix (i.e. a specific column) to each variable value of the same position (i.e. column) helping the subsequent modelling algorithm to focus on the variation between samples. It is usually recommended in combination with other methods for pre-processing of spectral data. The raw data arising from samples of "Vitamin A" (employed as raw material) and the outcome after the mean centring is shown in Figure 3-2a and b, where the mean on each column of the mean-centred data equals zero.

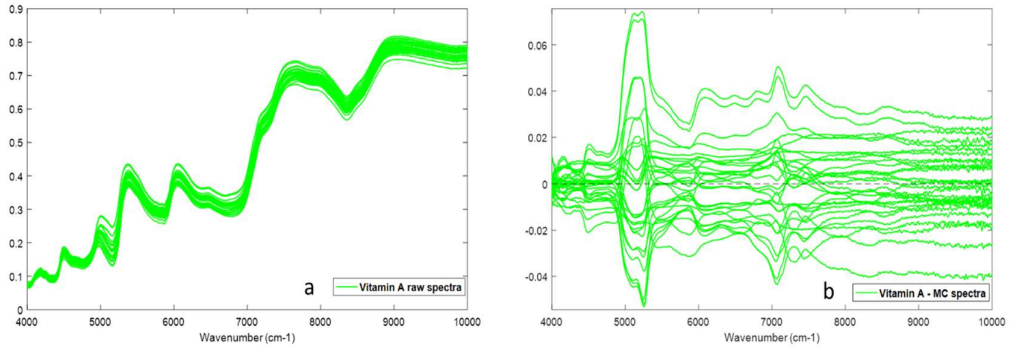

Figure 3-2. Raw spectral data (a) and mean centring (b) pre-processing outcome starting from the raw data of "Vitamin A".

- *Scaling: Unit variance scaling* divides each variable at a specific position by its corresponding standard deviation and is employed to make variables of different scales comparable prior to further processing. It is not applied to spectroscopic data as the relevant information is related to the differences in signal intensities. The integration of *unit variance scaling* with *mean centring* results in the so-called "*auto scaling"*.

The abovementioned methods belong to the "column-wise" treatments whereas the following methods refer to the "row-wise" treatments (O'Donnell, 2014).

3.2.1.2 Signal correction methods

According to Rinnan (2009), the row-wise pre-processing methods used in NIR spectroscopy can usually be grouped in scatter-correction methods and spectral derivatives. The class of the scatter-correction techniques consists of Multiplicative Signal Correction (MSC), Extended MSC (EMSC), Standard Normal Variate (SNV), Normalization, etc. The most used pre-processing techniques based on derivatives (first derivative that corrects the added baseline and the second derivative which corrects both offset and baseline slope), that use a first step of smoothing to reduce the noise arising from derivatives, belong to Savitzky-Golay (SG) polynomial derivative filters and Norris-Williams (NW) derivatives.

- *MSC*: The aim of Multiplicative Signal Correction is to remove the undesired scatter effects from the NIR spectral data before the subsequent modelling. It is one of the most applied scatter correction methods. The MSC algorithm includes two steps:

  - The first step ends up with the estimation of the parameters (a and b) that contribute to the additive and multiplicative effects (Equation 3.1), by means of least squares fitting:

$$\boldsymbol{Xorg} = \boldsymbol{a} + \boldsymbol{b} * \boldsymbol{Xref} + \boldsymbol{e} \qquad \text{Equation 3.1}$$

where $Xorg$ represents the original NIR spectra, $Xref$ refers to the reference spectrum that can be the mean spectrum of the training set used for pre-processing of the whole set of data, while e is the error (the un-modelled portion of the original spectra).

- The second step results in the correction of the obtained spectra (Equation 3.2).

$$\boldsymbol{Xcorr} = \frac{\boldsymbol{Xorg} - \boldsymbol{a}}{\boldsymbol{b}} \qquad \text{Equation 3.2}$$

where $Xcorr$ represents the corrected spectra.

An example depicting the raw spectral data (Figure 3-3a) of "Xangold" (raw vegetable material qualitatively assessed in Section 6 of this thesis, source of carotenoid esters) and the outcome of pre-processing by MSC is displayed in Figure 3-3b. The spectral response is expressed as reflected radiation.

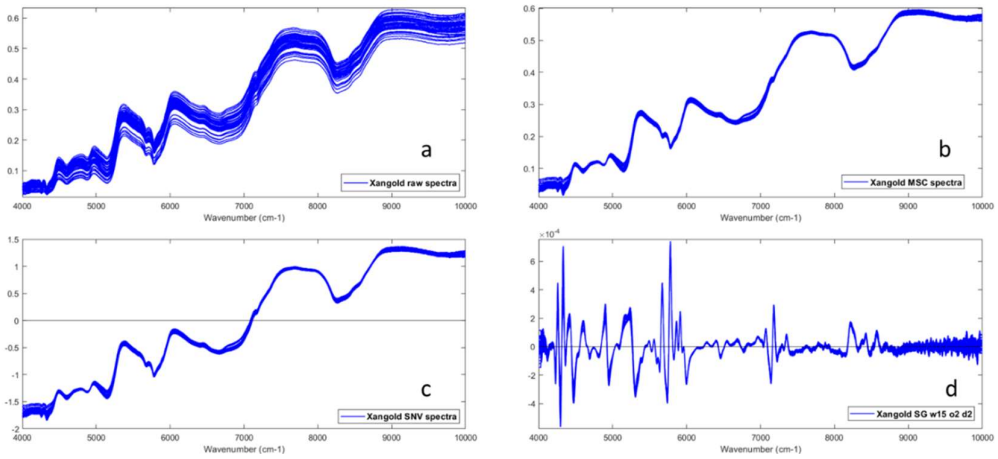

Figure 3-3. Raw spectral data (a), MSC (b), SNV (c) and Savitzky-Golay second derivative (d) pre-processing outcomes starting from the raw spectra of "Xangold".

While in MSC the correction is performed through a first-order polynomial, in EMSC a second order polynomial for reference correction is used, and the correction terms (parameters) are found by fitting a second order polynomial to the original spectra.

- *SNV*: There are analogies between SNV and MSC and the result in most practical cases is almost the same. The correction according to SNV is

operated on each spectrum individually, and it can be computed as (Equation 3.3):

$$\boldsymbol{X}corr = \frac{\boldsymbol{X}org - \boldsymbol{a}_0}{\boldsymbol{b}_0} \qquad \text{Equation 3.3}$$

where a_0 is the mean value of the NIR spectrum which has to be corrected and b_0 represents the standard deviation of the spectrum. The result of SNV pre-processing is shown in Figure 3-3c.
When the collected NIR spectra are noisy, as it can happen in on-line or in-line applications, a method named "robust normal variate" is recommended. This method employs as estimates for a_0 and b_0 the mean and the standard deviation of the inner quartile range, respectively. With regard to the normalization method, a_0 is set to zero and b_0 can be taken as the Euclidean norm or the normalization can be done over the variable with maximum absorbance or on a chosen wavelength.

- *Savitzky-Golay derivation*: By fitting a polynomial (the window size and the degree of the polynomial need to be selected) on the raw data, the parameters of the polynomial are computed, and the calculated derivative of the polynomial function is employed as derivative estimate for each data point. This procedure is applied to all the spectral data points with the order of derivative that depends on the polynomial degree. The outcome of this kind of pre-processing computing a second order derivative on the second degree polynomial trend using a window size of 15 points is displayed in Figure 3-3d.

- *Norris-Williams (NW) derivation*: The NW pre-processing involves at first a smoothing step of the spectral data, accomplished through the average over a selected number of points in the smoothing window. Then, the corrected spectrum is obtained by derivation through finite differences after setting a gap size between the smoothed values.

3.2.2 Principal component analysis (PCA)

PCA aims at projecting a set of samples, originally located in a high-dimensional space whose dimensions depend on the number of variables, onto a new space of lower dimensions, described by so-called principal components (PCs). The PCs represent the axes of the new space and the coordinates of each sample of the set are named "scores". In the new space, the main axis called first PC has the direction of the maximum variance of the samples and the second axis which is orthogonal to the main axis, depicts the maximum of the remaining samples dispersion. The following new axes are searched in this way until their number matches the number of original variables. By projecting the sample set on the new

PCs axes, the distances and the scales between them are preserved. PCA operates a reduction in dimensionality, making it possible to easily visualize the set of samples by means of two- or three-dimensional representations, obtained by plotting the scores one against the other (O'Donnell, 2014).

The data set is usually arranged in a matrix X where the number of rows represents the number of observations (samples) and the number of columns describes the number of variables. With respect to the spectroscopic data, the variables correspond to the spectral wavelengths. The application of a PCA algorithm decomposes the variance-covariance matrix of X into three matrixes i.e. the matrix T containing the scores, the matrix P containing the loadings (coefficients of the linear combinations of the original variables, each set corresponding to one principal component) and the matrix E containing the un-modelled part of the data (the residuals) with the same dimension as the original matrix (Figure 3-4). For a chosen number of components *k*, the matrix X having *n* rows and *p* columns can be decomposed as follows (Equation 3.4):

$$\boldsymbol{X}_{n\times p} = \boldsymbol{T}_{n\times k}\boldsymbol{P}^{\boldsymbol{T}}_{k\times p} + \boldsymbol{E}_{n\times p} \qquad \text{Equation 3.4}$$

where the number of columns of the scores matrix T equals the number of rows of the transposed loading matrix P^T. Since a total of *p* PCs can be extracted, the last *p–k* components are included in the error matrix E (O'Donnell, 2014).

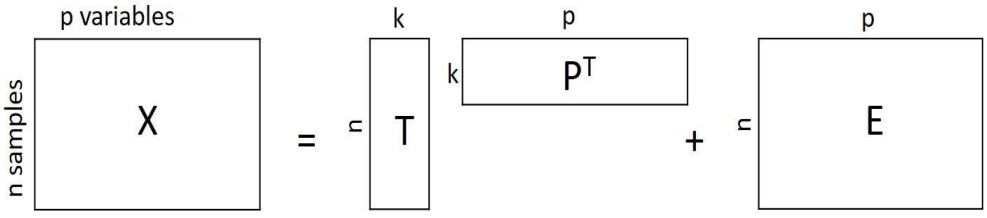

Figure 3-4. Visualization of the PCA decomposition structure.

When using continuous data like spectra, it is always recommended to pre-process the data by using the mean-centring step along with other techniques before applying the PCA algorithm. The loadings (principal components), which describe the explained variance in a decreasing order in all the samples, will be the same for all the averaged data. The score values, on the other hand, indicate how much each loading match the individual spectra and are different depending on the averaged spectra. The un-modelled part of the spectral data will be represented by the residuals E. If two samples have similar scores, it means they have similar chemical and physical characteristics (Bro, 2014). The relation between spectral variables and their contribution to the first and second PC as they explain most of the variance

is displayed in Figure 3-5a for a set of NIR spectral data arising from various lots of xangold.

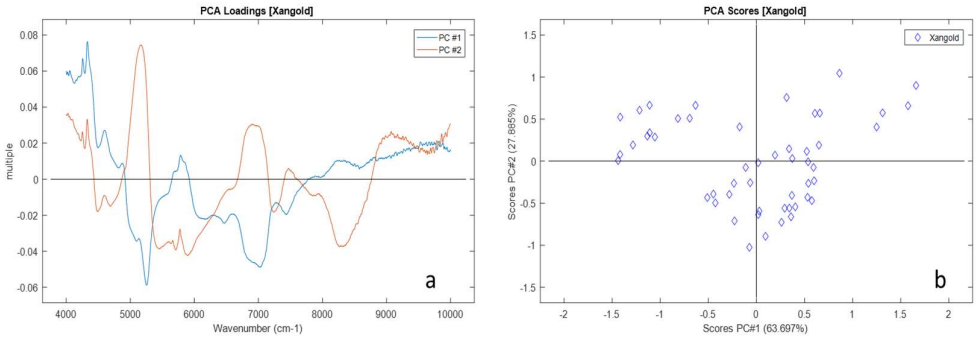

Figure 3-5. Loadings plot (a) and scores plot (b) of the raw material xangold, as obtained from the first and second principal components.

The relationship between the several lots (samples) of the abovementioned raw material is shown by the scores plot in Figure 3-5b where the first PC explains around 64% of the variance of the spectral data and the second PC around 28% for a total of 92% of variance explained by just two principal components.

3.2.3 Outlier detection

Within a set of samples subjected to multivariate data analysis there may exist outliers that are different from the other samples and as a result can un-properly perturb the model. Once the outliers are detected it is necessary to remove them from the analysis or keep them depending on why these samples are outliers (e.g. wrong labelling, measurement error, etc.) and how many within the data set are determined as such (if removing too many samples the model arising from the remaining samples may not be representative) (Bro, 2014). Outliers can be detected by means of a series of information and coefficients that results from the PCA decomposition such as:

Scores plot. One way to detect the outliers is by looking for atypical samples in the score plot of a set of NIR spectral data. It is important to investigate whether a sample can be considered an outlier or not looking at all the score plots according to the number of components. The scores plot in Figure 3-6 shows that the blue samples (ginkgo biloba dry extract 6%) are clearly different from the other plotted samples (ginkgo biloba dry extract 24%): these three samples can thus be identified as outliers.

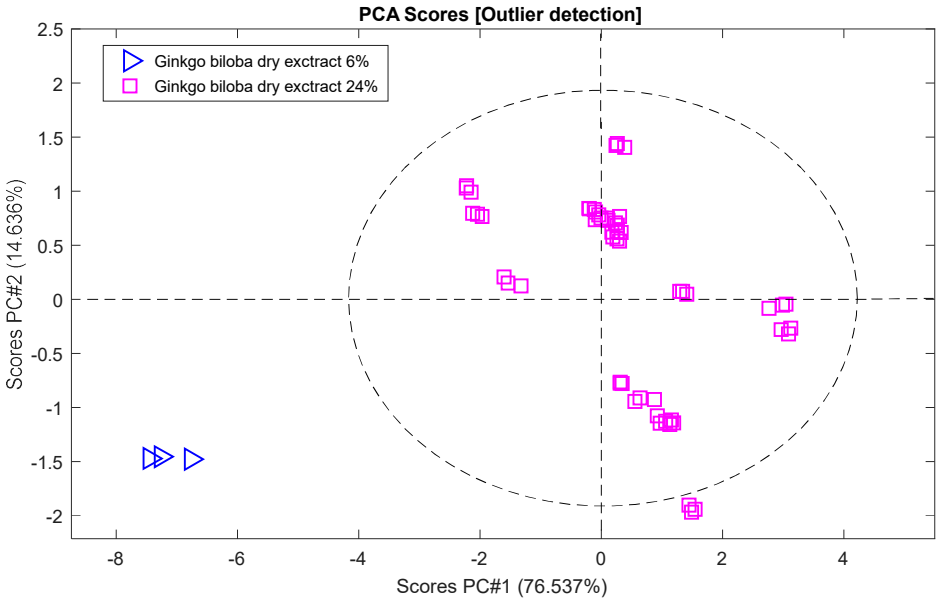

Figure 3-6. Outlier (ginkgo biloba dry extract 6%) detection by using scores plot and 95% of confidence limit.

Concerning the regression techniques such as PLS, which will be subsequently explained, the detection of outliers can be carried out by plotting the scores of the matrix Y against the scores of the matrix X.

Leverage or Hotelling's T^2. The most used way for the detection of outliers is through the leverage (h) or Hotelling's T^2 which are two tools that generally provide the same result related to the scores. The leverage computes the distance of each i-th sample from the centre of the model, according to the following equation (Equation 3.5), and is analogous to Mahalanobis distance (De Maesschalck, 2000).

$$h_i = \boldsymbol{t}_i(\boldsymbol{T}_0^T\boldsymbol{T}_0)^{-1}\boldsymbol{t}_i^T \qquad \text{Equation 3.5}$$

where T_0 corresponds to the matrix of scores and t_i represents the score vector related to the i-th sample.

The outliers that provide high leverage values ("strong" outliers) need to be explained as they have a huge effect on the model (O'Donnell, 2014).
Hotelling's T^2 can be employed in process monitoring (in-line or on-line) for the automatic detection of outliers comparing the scores of a given sample with the variation of the remaining data set, even if the visual representation of the model is a better approach (Bro, 2014).

Residuals or Q. Another tool used for outlier detection concerns residuals or Q distance. Residuals incorporate the un-modelled part of the spectral variance taking into account spectroscopic data. The Q distance refers to the distance of each sample

to the model's hyperplane. Outliers with a high residual variance are called "weak" outliers. The simultaneous detection of all types of outliers is achievable by plotting the residual variance against the T^2 (O'Donnell, 2014). This plot is named influence plot and the Figure 3-7 displays this plot considering NIR spectral data of Ginkgo biloba dry extract 24% and 6%. From this figure plot we can see that PC1scores capture 76.537% of the total variance whereas the remaining is residual variance. The plot shows that there are three large score outliers which represent the three spectra of the same sample of Ginkgo biloba 6%, and three slight residual outliers within the group of Ginkgo biloba 24%. The three samples of Ginkgo biloba 6% can be considered as outliers and removed from the plot.

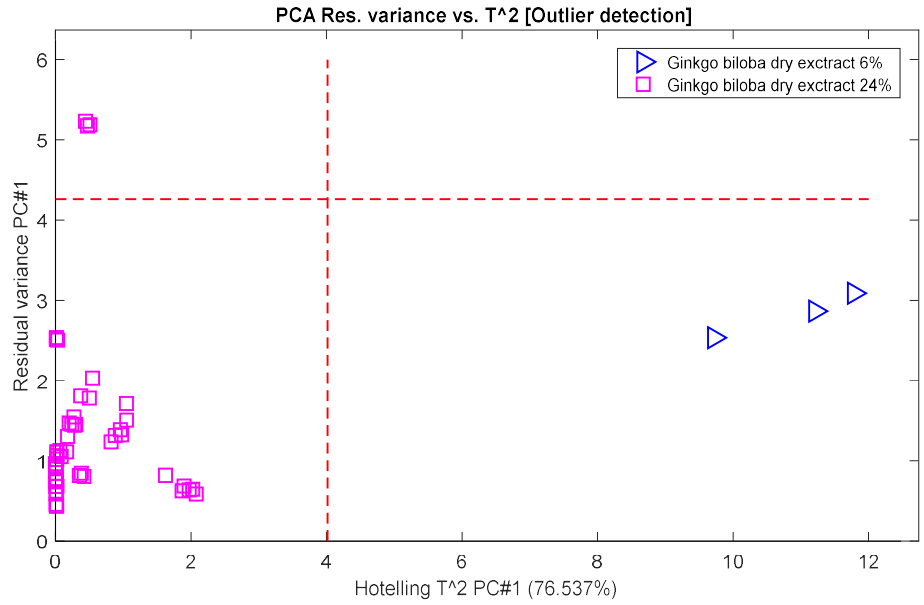

Figure 3-7. Influence plot of spectral data representing two Ginkgo biloba dry extracts according to a confidence limit of 95%.

3.3 Classification techniques

In univariate classification just one variable output is employed to classify objects whereas in multivariate classification several variables of the same sample are used for discrimination. Classification techniques are divided in "unsupervised methods", like clustering which aim to group samples searching for similarities, and "supervised methods" whose objective is to allocate samples to previously established classes (O'Donnell, 2014).

Unsupervised techniques consist of hierarchical clustering methods (all the samples are considered at the beginning) and non-hierarchical clustering methods, such as *K*-means, where a number of samples is left out at the beginning (O'Donnell, 2014).

Some of the supervised classification methods are described as follows.

3.3.1 Soft Independent Modelling of Class Analogy (SIMCA)

SIMCA belongs to the supervised classification techniques: each class is modelled independently using a specific PCA (the number of principal components used to build the various classes can be different). Usually, the limits of the classes are established through the confidence interval of the integration of Q distance and Hotelling's T^2 criteria or by computing the Euclidean distance of the residuals. SIMCA is referred to as a "soft modelling" technique because it allows the presence of regions in which two or more classes can overlap and hence a sample can belong to more than one class or none of the classes. On the contrary, with a "hard modelling" method discrete classes have to be build, and each sample is assigned to one class only. Four regions can be distinguished by plotting the residuals of two classes, built by applying the PCA algorithm, taking into account the confidence limits. Two zones are assigned respectively to the two classes, one zone representing the overlap between classes and the last zone within which fall the samples belonging to neither the classes (O'Donnell, 2014).

3.3.2 Partial Least-Square Discriminant Analysis (PLS-DA)

PLS-DA is an extension of the quantitative regression method (PLS) employed for qualitative analysis. The purpose of the PLS is to correlate the information between two blocks (usually multivariate) X and Y. In order to exploit the PLS benefits for classification purposes the matrix $Y(n \times m)$, which is called "dummy matrix", describes the m classes (represented by discrete numbers) included in the set of calibration samples. The two most important equations that explain the PLS-DA are shown below (Equations 3.6 and 3.7):

$$\boldsymbol{X = TP + E} \qquad \text{Equation 3.6}$$

$$\boldsymbol{Y = Tq + f} \qquad \text{Equation 3.7}$$

with T that represents the score matrix, P and q are the loadings, E and f represent the residuals (Brereton, 2014).

Figure 3-8 illustrates the model for a matrix X with n rows and p columns of which the n samples are divided in two equal classes (A and B) labelled with the binary code 1 and 0.

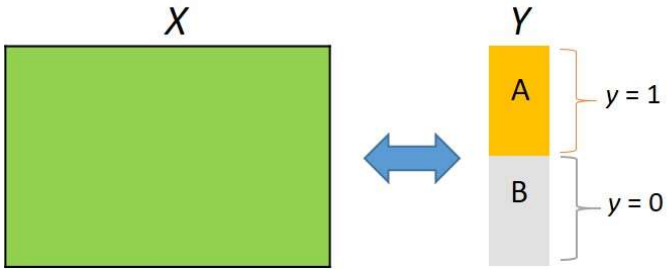

Figure 3-8. PLS-DA model of two classes depicted by the *Y* column vector (modified from: Brereton, 2014).

By applying PLS-DA, the matrix *X* (which contains the spectral data) will be described by a set of latent variables (LVs), which can be regarded as the linear combination of the original variables (similarly to the way principal components of PCA are built). The optimal number of LVs is decided using a cross-validation-based approach during the model training procedure (Ballabio, 2013).

3.4 Calibration techniques

In some specific situations the goal is to predict the quantity of one (y) or more (Y) quality attributes of interest. This aim can be achieved by correlating the attributes to a set of variables (X). These variables can be spectral data when spectroscopic techniques are used to collect the information on samples. The general equation has the following structure (Equation 3.8):

$$\boldsymbol{y_{pred}} = \boldsymbol{f(X)} \qquad \text{Equation 3.8}$$

where y_{pred} is the vector containing the predicted quantities of the attribute and *f* is the correlation function linking the spectral data of samples (X) and the quantity y to be predicted (O'Donnell, 2014). This function minimises the error between the actual responses (y_{ref}) and the predicted values (y_{pred}).

The regression model is first developed (calibrated) and optimized starting from the two known sets of data X and y and then used to predict the quantity attribute of unknown samples. Linear or non-linear regression methods can be used depending on the relation between X and y. Concerning the linear modelling, the following equation (Equation 3.9) combining X and y when X is multivariate (e.g. spectral data), contains the vector (b) of the regression coefficients of variables and the offset b_0 (O'Donnell, 2014):

$$\boldsymbol{y_{pred}} = \boldsymbol{b}_0 + \boldsymbol{Xb} \qquad \text{Equation 3.9}$$

3.4.1 Multiple linear regression (MLR)

MLR can be used when the number of variables in the original matrix does not exceed the number of samples and at the same time there is no collinearity between the variables. This method is not suitable for NIR spectral data as the number of variables is too high. The regression vector for this method is calculated on the original data (O'Donnell, 2014).

3.4.2 Principal component regression (PCR)

PCR differs from PLS as the scores of the only X matrix are calculated by the PCA algorithm. Then the regression vector (b) is calculated on the scores matrix T to avoid the effect of the possible correlation among the variables of the original matrix. The occurrence that the scores and loadings (PCs) are calculated only on the X matrix not taking into account of the Y matrix represents a drawback for PCR method (O'Donnell, 2014).

3.4.3 Partial least squares (PLS)

PLS regression is the most used linear multivariate regression method to predict the quantity of constituents of interest in samples. It correlates the information present in two different data matrixes. Considering a multivariate matrix X of spectroscopic data with *n* rows (samples) and *p* columns (variables) and a matrix Y with $n \times r$ dimensions containing the real values of the components. When the matrix Y contains only one variable the method is called PLS1 regression and when it is multivariate (two or more variables) it is called PLS2. A PCA algorithm decomposes the Y matrix in $\mathbf{Y} = \mathbf{U}\mathbf{C}^T + \mathbf{F}$ (U(scores), C(loadings) and F(residuals)) and the matrix X in $\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$ (T(scores), P(loadings) and E(residuals)). The values of Y can be predicted by predicting its scores (U). The PLS algorithm maximizes the covariance between T and U, i.e. the first score of Y has the maximum covariance with the first score of X ($\mathbf{u_1} = r_1\mathbf{t_1}$, and so on). In this way the prediction of the components of interest in an unknown sample can be made after collecting the relevant spectrum and knowing its score values which allow to predict the score values in Y and so to predict Y (Wold, 1983). An illustration of the PLS regression model is depicted in Figure 3-9.

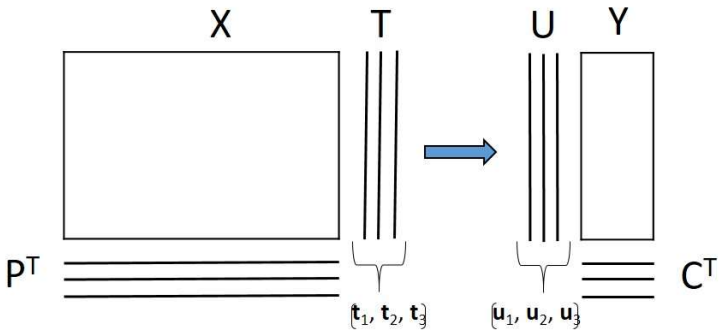

Figure 3-9. Graphical visualization of the PLS regression for a three component model (modified from: Wold, 2001).

3.5 Validation techniques

The developed models can be affected by overfitting or underfitting. The overfitting can arise from the high number of the latent variables employed in the calibration stage which may include in the model portions of the non-relevant part of the data. An overfitted model yields optimal but unrealistic results in calibration, but bad performance in the prediction of new samples. On the contrary, an underfitted model represents just a fraction of the data variability yielding poor result in calibration and prediction. In order to optimise the model avoiding the overfitting and underfitting, the model validation is required. Basically, there are two families of methods used for data validation to assess the quality of predictions: Cross-validation and test set validation (O'Donnell, 2014).

3.5.1 Cross-validation

The cross-validation method is the most used validation tool, based on an internal validation set. From the data set, one or more samples are removed at one iteration time. The remaining samples of the data set are used to develop the model, which is then employed to predict the removed sample or group of samples. All the samples are subjected to the abovementioned procedure in an iterative procedure, in which each sample will be removed at least one time. The leave-one-out is the less performing (less realistic) cross validation scheme where only one sample is removed at a time. The other methods remove blocks of samples in a continuous or random fashion. Cross-validation is usually used when the data set consists of a small number of samples (Brereton, 2007).

3.5.2 Test set validation

A better realistic result can be achieved when an external validation set is used. This implies the partition of the data set into two independent blocks which can be of different size. One of them is used as a training set for the model development whereas the samples of the other block called "test set" are predicted through the model (Brereton, 2007). Moreover, another data set made, for instance, of spectral data collected later than the first data set can be employed as external validation set.

3.5.3 Bootstrap

The bootstrap method has been proposed as a half way between cross-validation and the single test set. The test sets are internally generated in an iterative way and incorporate different combinations of samples where some samples can be part of various test sets. A prediction of the iteratively produced and afterwards removed test sets is performed accordingly (Brereton, 2007).

3.6 Model evaluation criteria

The most used criteria for the estimation of the predictive capacity of a developed model are the root mean squared error (RMSE) and the coefficient of determination (R^2). RMSE is computed by the root square of the ratio between the prediction error sum of squares, which is a measure of the closeness between the true (y_{ref}) and the predicted values (y_{pred}), and the number of samples (n) as shown in Equation 3.10. RMSE of calibration (RMSEC) is computed from the training set of samples. RMSE of cross-validation (RMSECV) is computed from the cross-validated samples whereas the RMSE of test set validation (RMSEP, prediction) is computed from the test set of samples (Nørgaard, 2000).

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum\left(\boldsymbol{y}_{ref} - \boldsymbol{y}_{pred}\right)^2}{\mathrm{n}}} \qquad \text{Equation 3.10}$$

The quality of the predictions is determined by the R^2 (Equation 3.11), that approaches 1 as the predicted values approach the true values.

$$\boldsymbol{R^2} = \mathbf{1} - \frac{\sum\left(\boldsymbol{y}_{ref} - \boldsymbol{y}_{pred}\right)^2}{\sum\left(\boldsymbol{y}_{ref} - \bar{\boldsymbol{y}}\right)^2} \qquad \text{Equation 3.11}$$

RMSEs and R^2 are both computed on calibration and validation sets. The plot reporting RMSEs versus the number of latent variables is then used to select the optimum number of components, which coincide with the minimum error in

validation. In order to avoid overfitting, the RMSEs of calibration and validation should be as close as possible (O'Donnell, 2014).

3.7 Process Monitoring and Control

Process improvement in industries can be achieved by variability reduction through statistical process control, which implies first the detection of the assignable causes (e.g. using control charts) and subsequently their removal. The variability can also be reduced taking advantage of the process regulation approach where the variability of the process output related to an established target can be minimized by regulating the process variables (Montgomery, 2009).

3.7.1 Multivariate Statistical Process Control (MSPC)

With respect to the production process there can be available different kind of sets of data: 1) prior the processing, 2) during the process and 3) after the processing (end product data). These data are important for process analysis, monitoring and control when process variables and quality attributes exceed their limits indicating that the process is out of the statistical control. Previously available data related to a given production process to be carried out is a starting point for understanding and monitoring purposes. The data belonging to the in-control processes or batches are used as reference for the future step. Having the data set of process variables (**X**) and quality attributes (**Y**) of the end product, a PLS regression can be performed (O'Donnell, 2014).

In a continuous process the **X** matrix is bi-dimensional (time points × number of variables), whereas when working with batches a tri-dimensional matrix is generated with each row representing a batch. The score and the loading plots, related to the process variables, can differentiate the processes or batches in clusters and provide the information to explain them.

The quality assurance of a product can be achieved through process monitoring and control instead of relying only on the final product measurements. Process monitoring allows the real time assessment of the process based on the decided control limits. Corrective actions need to be taken when the process is out of specifications depending on the nature of deviation.

Univariate control charts like the Shewhart (Brereton, 2007) can be used for monitoring purposes. These charts allow the monitoring of different process parameters (e.g. quality characteristics) independently. The evolution (systematic or random) of single parameters provides information about the process. When these characteristics exceed their upper or lower control limit, actions must be taken. However, inspecting the variables independently can result in misleading considerations. In order to get more accurate results, it is recommended the use of multivariate control charts, which are based on the covariance between variables. The Hotelling's T^2 multivariate control chart is based on computing the multivariate distance between the sample and the centre of the model (built on reference data).

Another multivariate control chart called squared prediction error (SPE) is based on the residual variance of each sample. The instruments used to understand the cause of a deviation from the limits which arise from loadings are called contribution plots representing each sample and control chart (O'Donnell, 2014).

Process monitoring, sometimes, depends on the spectroscopic data like NIR spectra, where each spectrum is acquired in a few seconds. Building a control chart for each variable would become difficult to manage as the number of variables is very high. One possible solution is to use multivariate qualitative methods based on PCA to represent the samples on a bi-dimensional space. PLS regression methods allow to monitor the quantity of one or more response variables with the process time, starting from previous calibrations arising from NIR data and the quantitative variables of interest (Brereton, 2007).

The monitoring of the batch evolution is accomplished through the batch statistical process control by comparison with a reference batch. The real time batch monitoring and the later release is performed by methods based on multi-way PCA and multi-way PLS which, differently than PCA and PLS that work on bi-dimensional matrices of data, are tools that performs PCA and PLS in multidimensional arrays as it happens when different batches of a process are considered at the same time. The corrective actions undertaken, arising from a deviation during the process, belong to the feed-back control procedure. Moreover, a control procedure arising from a model built on the current measured quality of the raw materials is part of the feed-forward control (Kourti, 2006). A process can therefore be controlled both looking at (monitoring) its outputs (feed-back) and its incoming raw materials which can change from batch to batch or lot to lot (feed-forward).

3.8 References

http://www.gruppochemiometria.it/

Montgomery, D. C. (2017). *Design and analysis of experiments*. John wiley & sons.

Leardi, R. (2009). Experimental design in chemistry: a tutorial. *Analytica chimica acta*, *652*(1-2), 161-172.

Brereton, R. G. (2007). *Applied chemometrics for scientists*. John Wiley & Sons.

Rinnan, Å., Van Den Berg, F., & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, *28*(10), 1201-1222.

O'Donnell, C. P., Fagan, C., & Cullen, P. J. (Eds.). (2014). *Process analytical technology for the food industry*. Springer.

Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical Methods*, *6*(9), 2812-2831.

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, *50*(1), 1-18.

Brereton, R. G., & Lloyd, G. R. (2014). Partial least squares discriminant analysis: taking the magic away. *Journal of Chemometrics*, *28*(4), 213-225.

Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods*, *5*(16), 3790-3798.

Wold, S., Martens, H., & Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. In *Matrix pencils* (pp. 286-293). Springer, Berlin, Heidelberg.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, *58*(2), 109-130.

Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J. P., Munck, L., & Engelsen, S. B. (2000). Interval Partial Least-Squares Regression (i PLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. *Applied Spectroscopy*, *54*(3), 413-419.

Montgomery, D. C. (2009). *Statistical quality control* (Vol. 7). New York: Wiley.

Kourti, T. (2006). Process analytical technology beyond real-time analyzers: The role of multivariate analysis. *Critical reviews in analytical chemistry*, *36*(3-4), 257-278.

Chapter 4

Process Analytical Technology and Quality by Design

Process analytical technology (PAT) term was introduced by the US Food and Drug Administration (FDA) in 2004 and defined as: "A system for designing, analyzing, and controlling manufacturing through timely measurements (i.e., during processing) of critical quality and performance attributes of raw and in-process materials and processes with the goal of ensuring final product quality" (FDA, 2004). The PAT has its origins in the process analytical chemistry (PAC) developed in the early part of the 20th century. PAC strategy was developed and applied mainly for the chemical industry with the aim at monitoring the related processes on an at-line at the beginning and later on/in-line, real-time based approach. In order to extend the field of application to other industries such as pharmaceutical, biochemical and so on the term "technology" was employed instead of "chemistry" (Workman, 2011). PAT is recognized as an important tool for the monitoring and control of manufacturing processes of pharmaceutical and food industries and has become an essential part of the more recent and comprehensive approach known as Quality by Design (QbD). In the following paragraphs, after a brief introduction to the QbD concept, the PAT approach will be further detailed.

4.1 Quality by Design (QbD)

A broader approach which includes the PAT tool, known as Quality by Design (QbD), first developed by Joseph M. Juran (Juran, 1992), was defined in 2009 as: "A systematic approach to development that begins with predefined objectives and emphasizes product and process understanding and process control, based on sound science and quality risk management" (ICH Q8 (R2), 2009). The elements making up the QbD approach with respect to the product development within the pharmaceutical industry can include:

- The identification of the critical quality attributes (output product properties that should comply with limits) of the product by the quality target product profile (QTPP) which represents the starting point for the design of product development. QTPP may depend on the intended use, dosage form, delivery structure, etc.

- Product design and understanding, that aims at developing a product that, guaranties its expected performance during its shelf life and, meets the final user's needs validated through stability and clinical studies. The successful design and development of a product should be based on the characterization (chemical, physical and biological) of active ingredients, selection of other ingredients (excipients) taking into account their natural variability, interaction of several ingredients, identification of critical material attributes (input material properties that should comply with quality limits) of the ingredients and in-process materials, and at last but very important the formulation optimization.
- Process design and understanding. The critical quality attributes (CQAs) of an end-product depend on the critical material attributes (CMAs) and the critical process parameters (CPPs, parameters which need to be monitored or controlled as their variability affects the critical quality attributes). In general, process robustness studies are performed to assess the outcome (in terms of product quality and performance) of process parameters and material attributes changes. According to the ICH Q8 (R2), it is also defined the design space as: "The multidimensional combination and interaction of input variables (e.g., material attributes) and process parameters that have been demonstrated to provide assurance of quality". A process understanding is achieved through the identification of the critical sources of variability and the accurate prediction of quality attributes of the product, managing the variability by the process. Usually, the manufacturing process involves several steps to get the desired final product and the identification of the critical attributes and parameters may be performed on every unit operation or a combination of unit operations following the steps of process understanding.
- Control strategy. Control strategy is defined by the guideline ICH Q8 (R2) as: "A planned set of controls, derived from current product and process understanding that ensures process performance and product quality". There can be three levels of control. The first level of control (level 1), which is based on automatic engineering control, guaranties the final product quality through the monitoring of the attributes of incoming materials and the real-time adjustment of process parameters. Real time release testing can be achieved by this control level employing the PAT tool. The second level of control (level 2) includes a limited number of analysis on the final product and flexibility related to the input material attributes and process parameters into the design space. The third level of control (traditional level) consists in large final product analysis and restricted understanding of material attributes and process parameters. Control strategy as an integrated approach along the production chain can be based on the control of incoming material attributes, product specifications, controls related to unit operations that can affect the following processing or product quality, analysis and control of critical quality attributes over processing and a monitoring program at established

periods of time based on product analysis with the aim of assessing the accuracy of multivariate prediction models.

- Process capability and continual improvement. A process subjected to discontinuous variation (special cause) doesn't result in a state of statistical control and the related measurements lead to process performance index. Besides the special causes, the process can be affected by random, regularly present common causes that are responsible of intrinsic variation. An estimation of the process capability can be done starting from the specification limits and the intrinsic variability related to a stable process in a state of statistical control. The potential sources of intrinsic variation can be identified in an early stage of product development and as a result keep them under control due to product and process understanding components along with the control strategy of the QbD approach. The absence of a QbD approach during manufacturing may bring to commercial production interruption when the intrinsic variability is detected during this step. The elimination of intrinsic variability sources from the process operation status and incoming raw materials can be achieved by continuous improvement with the aim of improving the process measurable by process capability. Multivariate analysis can be employed to investigate previous manufacturing data with the purpose of gaining knowledge on the variability of raw materials and process parameters. The reduction and control of the abovementioned variability will result in continuous improvement.

There also exists some risk associated with the manufacturing and use of a final product. The risk assessment is conducted to identify the critical variables, helping the control strategy development and implementation. Tools like risk ranking and filtering, hazard analysis and critical control points, etc. can be employed for risk assessment purposes. The implementation of QbD aim at reducing the product variability, improving manufacturing performances and product development where PAT is a tool which helps to achieve the goals of QbD (Yu, 2014). The principles of QbD can also be applied in food industry with the support of PAT technology (van den Berg, 2013).

4.2 Process Analytical Technology (PAT)

PAT can be used as a functional tool for good manufacturing practice (GMP) system due to its advantage in improving the production processes. Several processes in the food and pharmaceutical industry need continuous validation to guarantee their effectiveness and the PAT approach can facilitate the achievement of this objective. The systematic preventive approach known as Hazard Analysis and Critical Control Points (HACCP) is adopted in the food industry with the aim at reducing the risks associated to chemical, physical and biological hazards. The PAT technology can also be integrated in the HACCP system to assure food quality

and safety taking advantage of the process monitoring. The benefits of PAT application in production processes may concern environmental sustainability using more efficiently the resources and reducing costs and environmental impacts (O'Donnell, 2015).

The ever higher quality and safety standards of products considering the variability of raw materials and the variation in processes, can be achieved by moving from traditional discontinuous time consuming analysis to rapid methods of analysis within the PAT framework which yield a high number of analytical responses. The expected goals in both food and pharmaceutical industry are the same and can include: high product quality and safety, low use and consumption of natural and human resources, restrained effect of raw materials variability, increased productivity and shelf life of products. The food industry, particularly, depends on soft, fragile and not pure raw materials that consist on various compounds. Environmental and storage conditions as well as processing affect easily the physical properties of these raw materials. Moreover, the physical, biochemical and microbiological processes at the micro-scale could not be known (Hitzmann, 2015). Process monitoring of the food industry tends to be more difficult compared to monitoring of the pharmaceutical production processes as foods are complex matrixes (solids, liquids, gels, etc.) made of heterogeneous classes of compounds such as carbohydrates, fats and proteins along with the micronutrients (van den Berg, 2013).

4.2.1 PAT principles

The PAT implementation can enable manufactures to deliver a product, that meets the specifications, through the active process control procedure avoiding the post-process conformity analysis. This is a shift from the common feed-backward process control (post-problem), where the corrective actions on the process are taken after the final product analysis, to predictive process control (during problem) where the process adaptations over manufacturing compensate the high raw materials variability (Figure 4-1). The active process control is facilitated by the developments made in the measurement tools which allow to perform on/in-line analysis. The benefits arising from the effective implementation are: the effective use of raw materials, the reduced variation of quality attributes of the end product, the rework and/or waste reduction, the replacement of slow and expensive laboratory analysis, and continuous learning. The diffusion of the PAT is facilitated by the advances in spectroscopic monitoring techniques equipped with fibre optic probes and chemometrics (van den Berg, 2013).

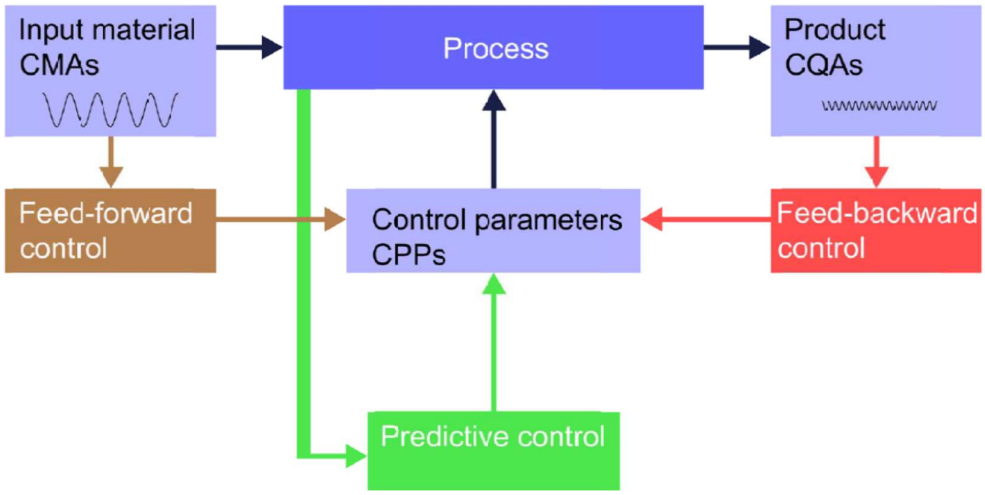

Figure 4-1. Schematic representation of the monitoring and control process involving a manufacturing unit operation (modified from: van den Berg, 2013).

The employment of PAT tools involves the prior considerations of some features which can include: 1. The efficacy of the process analyser to achieve the purposes of process monitoring or control and product information; 2. The process analyser location depends on the production process and the expected information to obtain; 3. The favourable measurement conditions of the process analyser in order to get data that represent the real process; 4. The performance of the process analysers should be validated at needful intervals of time. Moreover, other factors such as food matrix characteristics, the degree of process automation and the robustness of validated PAT sensors used for in-line applications have to be assessed (Panikuttira, 2018).

4.2.2 PAT components

According to Van den Berg (2013), the implementation of PAT relies mainly on four constituents, taking into account the developments in process testing technology, fibre optics and computational methods. A description of the abovementioned components is given below.

4.2.2.1 Critical product quality attributes and process parameters

One of the main steps in PAT implementation concerns the dependence of the critical quality attributes from the process operating parameters. Understanding this relation may be useful for establishing the allowable limits regarding the variation of quality attributes which supports the decision on the suitable sensor choice. On the other hand, the impact of process parameters on the critical quality attributes can be understood through the application of PAT.

4.2.2.2 Process dynamics and sampling

There are other aspects to take into account during the implementation of PAT such as: the features of process analyser (sensor), the efficacy of the sampling procedure to represent the process and the process dynamics. The relation among the three abovementioned factors of process understanding can be illustrated by five elements as follows:

- *Measurement uncertainty.* The robustness and sensitivity of a sensor employed in a given application rely on its measurement uncertainty (which depends on process environment) and the model built for prediction purposes. Moreover, the value of the sensor is assessed as the ratio of the measurement error and the natural process variability.
- *Frequency of measurements.* The better approximation of the real process signal is accomplished by the on/in-line sensors as they provide a high frequency of measurements compared to the at/off-line analytical techniques.
- *Time of measurement.* The utility of the sensor response depends also on the time needed for the measurement. In order to perform real time monitoring a short time of measurement is required.
- *Carry over effect between measurements.* The memory effects, which can result in delayed response, originated from some process sensors and sampling designs are the cause of carry over signals. Spectroscopic sensors equipped with fibre optic probes designed for non-invasive analysis can overcome this problem.
- *Lapse of time between sampling and measurement.* In order to perform real time process monitoring and control the off-line measurements are usually unsuitable as the process will change significantly during the time between sampling and analysis. By using spectroscopic techniques, the lapse of time tends to be reduced or eliminated.

4.2.2.3 Spectroscopic techniques and imaging

Spectroscopic techniques are key tools for the implementation of PAT due to their characteristics in providing multiple chemical information in a fast and non-destructive fashion. Besides the NIR spectroscopy illustrated in chapter 2, other spectroscopic techniques such as fluorescence and Raman spectroscopies, which are briefly described below along with ultraviolet-visual (UV-VIS) and infrared (IR) absorption spectroscopies, can be employed as tools to achieve the PAT objectives. Besides their high sensitivity, the drawback of UV-VIS and fluorescence sensors, which involve electronic transitions in substances, is related to the limited number of molecules that can be tested (van den Berg, 2013).

- Fluorescence spectroscopy

The phenomenon of fluorescence is related to the emission of energy following the transition of the previously exited molecule from the exited electronic state of singlet to the ground electronic state (Guilbault, 1990). The fluorescence spectroscopy has the advantage of being the most sensitive among the spectroscopic tools and can be used for the monitoring of raw materials, processes and product quality attributes. Most of the applications involving this technique up to now concern the laboratory scale instead of the industrial PAT applications. The monitoring of fermentation processes such as that of rye sourdough has been carried out coupling fluorescence with chemometric techniques like PLS and PCR in order to predict pH and acidity values (Grote, 2014). Another study aimed at improving the food security has been focused on the possible determination of acrylamide by converting it to a compound with a strong fluorescence emission (Liu, 2014). Based on previous results, some authors proposed the application of fluorescence imaging for routine testing in delicatessens (Beck, 2015). Some scholars proposed a method based on hyperspectral imaging and multivariate image analysis tools comparing fluorescence imaging arising from the visible region to that deriving from violet and ultraviolet spectral region (Hitzmann, 2015).

- Raman spectroscopy

The Raman spectrum is characterized by two set of lines called Stokes (greater intensities) and anti-Stokes arising from the inelastic collision between the incident monochromatic radiation (usually visible region at high wavelength) and the molecule (Colthup, 1975). Surface-enhanced Raman spectroscopy (SERS), instead, is based on the occurrence that the Raman scattering signal can be enhanced (chemical or electromagnetic enhancement) when the scatterer molecule is positioned on or close the surface of a roughened noble-metal substrate (Hynes, 2005). Some scholars have taken advantage of the benefits provided by SERS in combination with chemometric techniques to identify prohibited food additives for food safety purposes (He, 2015). The quantitative determination of soft drinks ingredients such as glucose, fructose and sucrose has been accomplished by using a rapid method based on Raman spectroscopy (Ilaslan, 2015). Through another study was demonstrated the capability of Raman technique as process sensor for on-line monitoring of the wine fermentation constituents such as sugar, ethanol and glycerol employing high pressure liquid chromatography (HPLC) as reference method (Wang, 2014). The combination of Raman spectroscopy with multivariate data analysis resulted also effective for on-line monitoring of the meat quality.

4.2.2.4. Chemometrics

The quality monitoring and control of industrial production processes has undergone a revolution due to developments in multivariate spectroscopic sensors (e.g. NIR spectroscopy equipped with optic fibres) and chemometrics. During a process' batch, the large amount of NIR spectral data, collected automatically by a

sensor which is connected to a production process, need first to be analysed and then the information obtained is transformed into real knowledge about the process through chemometric models. The analysis of information provided by spectroscopic sensors through multivariate bilinear exploratory methods like PCA and regression methods such as PLS brought to good demonstrated results for quality control. The benefits ensuing from the use of the abovementioned chemometric methods can bring to new process and production understanding, which can be used to monitor or control the critical sources of variability, as well as to more efficient calibration models due to their advantages in the exploration of the process trajectory, outlier detection, etc. (van den Berg, 2013).

4.3 NIR analysis strategies

According to the location of the analytical instrumentation, the ways to perform measurements can be distinguished in: off-line, at-line, on-line, in-line and non-invasive as illustrated in Figure 4-2.

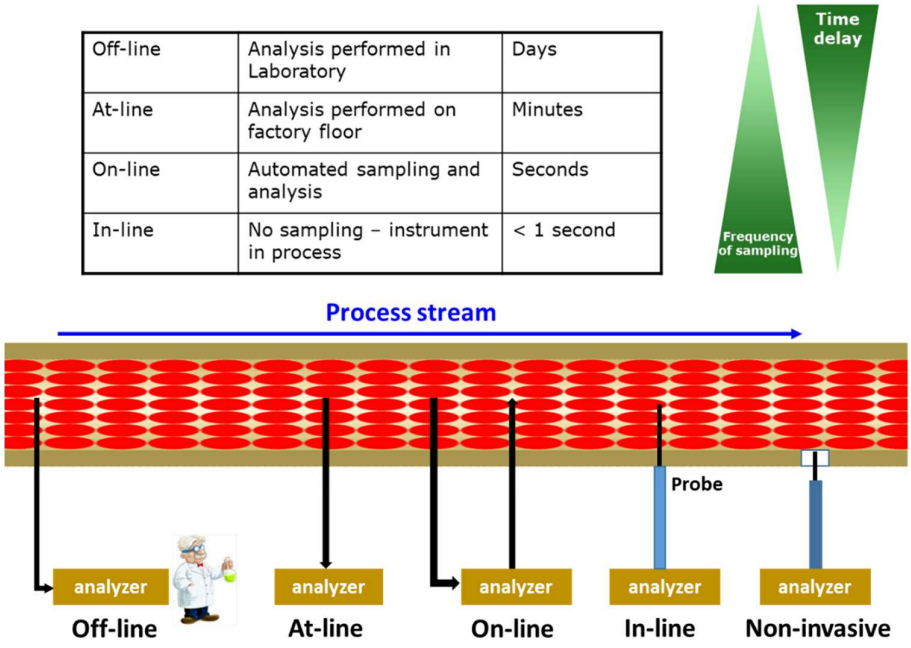

Figure 4-2. Representation of the several strategies which can be used to test the product throughout the process.

4.3.1 Off-line testing

For off-line measurement the sample, once removed manually from the process line, is then transported to a centralized support laboratory to perform the analysis. Besides the advantages related to the laboratory analysis in terms of the availability of an expert team, some disadvantages such as costs and the delay between

sampling and provided results are typical. This delay is critical especially when the purpose of testing is related to process monitoring and control. Therefore, this approach is usually used when discontinuous process information is required. However, NIR spectroscopy has the advantage to perform analysis in few seconds compared to time consuming methods like HPLC.

4.3.2 At-line testing

In order to perform at-line testing the analytical instrument is located within the production area in the proximity of the process line. Usually the instrumentation used is simpler and more robust than the benchtop laboratory instruments, suited for a rough environment and aimed at performing specific analysis on samples collected from several locations within the production facility, with higher ease of use and the dedicated process team can provide faster responses that can fit for process monitoring and control purposes.

4.3.3 On-line testing

On-line analysers systems enable automated sampling with injection of a little quantity of sample stream within the analyser (intermittent methods) or continuous measurements as the sample flows continuously though a by-pass loop (continuous methods). These systems can provide sample pre-treatments before analysis where the end point of the analysis concerns the critical quality attributes instead of the chemical composition generally provided by off-line analysers. External devices such as pumps can be controlled by on-line systems of testing. The instrumentation based on spectroscopic sensors is typically employed for continuous on-line process measurements providing real time analytical responses.

4.3.4 In-line testing

The in-line process instruments are equipped with a chemically sensitive probe (ideally it can be similar to a temperature probe) which is put directly into the process stream providing in-situ measurements (real time information) with no need to use a side stream as required for the on-line analysis. For example, the developments in fibre optic probes enabled the implementation of in-line analyses. In this equipment the light generated from a remote source is conveyed through a fibre at the distal end of the probe (e.g. reflection probe) which is immersed into the process stream. The reflected radiation, arising after the interaction with the sample, can be collected by another fibre and conveyed into the instrument (usually spectroscopic tools) for testing.

4.3.5 Non-invasive testing

This strategy of analysis yields the analytical response while the probe is not physically in contact with the sample, for example using a transparent window directly positioned on a pipe or a fermenter/reactor through which an optical fibre can convey and collect the light chosen for performing the analysis (Callis, 1987).

4.4 PAT implementation in food and pharmaceutical industries

The at-line monitoring of vegetable oil deterioration, made of two different oils, during the frying process has been performed by using spectroscopic sensors combined with chemometric methods (Engelsen, 1997). Different chemical and physical quality attributes, such as anisidine number, free fatty acids, vitamin E, etc., have been measured during the process. It is shown that the development and the complexity of the process is not addressed by monitoring one variable at a time. The PCA model build on the NIR spectra collected during the process explained the drying process evolution and process dynamics over time. This implementation is a form of real-time PAT application and had the aim to match the sensor characteristics to the process dynamics gaining knowledge (van den Berg, 2013).

An on-line system, based on NIR spectroscopy and chemometrics, for automatic process monitoring and control of ammonia dosing to produce low methoxylated amidated pectin from high methoxilated pectin has been successfully implemented (Zachariassen, 2005). This study covered all the aspects related to the PAT development. The implementation of a selected NIR sensor, performing high frequency measurements, revealed a considerable fluctuation of ammonia quantities after dosing which were not noticed employing the reference method. The acquired process knowledge about process dynamics has been used for real-time control of ammonia providing a reduction of process fluctuations. Accordingly, a series of benefits resulted such as product uniformity, compliance with specifications, scrap and environmental footprint reduction through the more controlled dosing of reactant and the less usage of traditional chemical testing (Zachariassen, 2005).

According to various studies there is a strong potential and need for the PAT adoption in cheese manufacturing with the aim at improving product quality and enhancing process efficiency (Panikuttira, 2018). PAT components such as spectroscopic sensors have the potential to achieve these objectives considering the recent technological advances of these tools. Milk quality parameters, which affect the cheese quality and yield, including the presence of adulterants and contaminants as well as the key unit operations (standardization, coagulation, syneresis and ripening) involved in cheese manufacturing will increasingly be monitored by NIR, MIR, fluorescence and Raman sensors in combination with chemometrics (Panikuttira, 2018).

The implementation of PAT is a step forward towards Industry 4.0 which will lead to enhancement of raw materials validation, process optimization, better traceability, increased quality and automation across the production due to the connectivity of factory components (Panikuttira, 2018).

The characterization of the alginate (binary polysaccharides) composition by a fast and at-line FT-IR method has been accomplished (Salomonsen, 2008). Alginates are obtained from the brown algae and used in food and pharmaceutical industry as thickeners, gelling agents and stabilizers. For the purpose of the study, a calibration data set of 100 alginates samples covering the whole ratio (parameter that affect alginate properties) range between the two compositional monomers was subjected to testing performed either by the reference, time consuming, off-line nuclear magnetic resonance method or by the FT-IR technique. The correlation coefficient and the prediction error arising from the multivariate PLS regression model have been considered acceptable for the application. In this implementation, the process understanding resulted from this post-production PAT employment can then be used for future product improvements (van den Berg, 2013).

A feed-forward PAT application showed its potential for the classification of porcine carcasses based on fat quality in abattoirs. To achieve this target an on-line spatially resolved NIR method was developed to predict the degree of fatty acid unsaturation as a function of the penetration depth, employing the Gas Chromatography technique as reference method (Sørensen, 2012).

A PAT approach based on the combination of FT-NIR and chemometrics demonstrated the potential for in-line process monitoring of hot melt extrusion (HME) (Vo, 2018). HME is a process that can also be used to create dosage forms for active pharmaceutical ingredients (API) with low solubility and in which the assurance of product quality and uniformity is achieved through monitoring and control of critical process parameters. On a chosen NIR spectral interval a robust PLS model has been developed for the quantitative prediction of a API. Moreover, the evolution of the process depending on the critical process parameters such as temperature and the rate of incoming material has been monitored by a PCA qualitative model (Vo, 2018). A PAT system based on UV/VIS spectrometer was implemented for in-line monitoring of the critical quality attributes (carbamazepine and theophylline) of a dosage form (solid dispersion) produced by hot melt extrusion where the active ingredients are distributed in a hydro-soluble polymer matrix (copovidone) carrier (Wesholowski, 2018).

Blending is one of the unit operations forming the pharmaceutical manufacturing process and the monitoring and control of the progress of this step is necessary to ensure product quality and identify the end point. The implementation of PAT to assess mixing uniformity and content can lead to benefits such as better process monitoring and increased manufacturing process understanding compared to the current strategy of blend uniformity assessment. NIR spectroscopy in combination with chemometric methods is the most investigated technique among the large number of PAT sensors studied for homogeneity purposes. However, for the implementation of these sensors to a particular system, the drawbacks such as the location and adaptation of the

equipment to provide real process information, costs, etc. should be kept into account (Crouter, 2019).

In the continuous manufacturing, a PAT strategy based on NIRS and multivariate data analysis demonstrated its usefulness for the in-line monitoring and control of the tableting step (Pauli, 2019). Two strategies consisting on NIRS and the PLS regression method were successfully developed and optimized, using HPLC as a reference technique, for the assessment of the API content uniformity in the mixture within the feed frame and for the monitoring of the tablet content uniformity at different tableting speeds. Both calibration models displayed outstanding predictive capabilities, for the active component, assessed by the correlation coefficients and the root mean square errors. The development and the implementation of calibration models for predicting the quantity of water in tablets is an important task as the water is a critical quality attribute which effects product shelf life (Pauli, 2019).

The development of a method based on NIRS for the quality control of 3D printed tablets of different geometries, excipients and concentrations of paracetamol has been validated showing an outstanding linearity and accuracy in prediction. Information about the dispersion and the crystalline/amorphous content of the active ingredient in the several dosage forms has been achieved by using Raman confocal microscopy (Trenfield, 2018).

4.5 References

U.S. Department of Health and Human Services, Food and Drug Administration, *Guidance for Industry, Process Analytical Technology – A Framework for Innovative Pharmaceutical Development, Manufacture and Quality Assurance*. Pharmaceutical CGMPs, September 2004.

Workman, J., Lavine, B., Chrisman, R., & Koch, M. (2011). Process analytical chemistry. Analytical Chemistry, 81, 4623-4643.

Juran JM. Juran on quality by design: the new steps for planning quality into goods and services. New York: The Free Press; 1992.

ICH Q8R2 Pharmaceutical Development. International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use, August 2009.

Lawrence, X. Y., Amidon, G., Khan, M. A., Hoag, S. W., Polli, J., Raju, G. K., & Woodcock, J. (2014). Understanding pharmaceutical quality by design. *The AAPS journal*, *16*(4), 771-783.

O'Donnell, C. P., Fagan, C., & Cullen, P. J. (Eds.). (2014). *Process analytical technology for the food industry*. Springer.

Hitzmann, B., Hauselmann, R., Niemoeller, A., Sangi, D., Traenkle, J. & Glassey, J. (2015). Process analytical technologies in food industry - challenges and benefits: a status report and recommendations. Biotechnology Journal, 10, 1095–1100.

van den Berg, F., Lyndgaard, C. B., Sørensen, K. M., & Engelsen, S. B. (2013). Process analytical technology in the food industry. Trends in Food Science and Technology, 31, 27–35.

Panikuttira, B., O'Shea, N., Tobin, J. T., Tiwari, B. K., & O'Donnell, C. P. (2018). Process analytical technology for cheese manufacture. International Journal of Food Science and Technology, 53, 1803−1815.

Guilbault, G. G. Practical Fluorescence; 2nd ed.; Dekker: New York, 1990

Grote, B., Zense, T., & Hitzmann, B. (2014). 2D-fluorescence and multivariate data analysis for monitoring of sourdough fermentation process. *Food Control*, *38*, 8-18.

Liu, C., Luo, F., Chen, D., Qiu, B., Tang, X., Ke, H., & Chen, X. (2014). Fluorescence determination of acrylamide in heat-processed foods. *Talanta*, *123*, 95-100.

Beck, E. A., Lefcourt, A. M., Lo, Y. M., & Kim, M. S. (2015). Use of a portable fluorescence imaging device to facilitate cleaning of deli slicers. *Food Control*, *51*, 256-262.N.B. Colthup, L.H. Daly, S.E. Wiberley, Introduction to Infrared and Raman Spectroscopy, 2nd ed. Academic Press, New York, 1975.

Haynes, C. L.; McFarland, A. D.; Van Duyne, R. P. Surface-Enhanced Raman Spectroscopy. Anal. Chem. 2005, 77, 338A–346A.

He, S., Xie, W., Zhang, W., Zhang, L., Wang, Y., Liu, X., ... & Du, C. (2015). Multivariate qualitative analysis of banned additives in food safety using surface enhanced Raman scattering spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *137*, 1092-1099.

Ilaslan, K., Boyaci, I. H., & Topcu, A. (2015). Rapid analysis of glucose, fructose and sucrose contents of commercial soft drinks using Raman spectroscopy. *Food Control*, *48*, 56-61.

Wang, Q., Li, Z., Ma, Z., & Liang, L. (2014). Real time monitoring of multiple components in wine fermentation using an on-line auto-calibration Raman spectroscopy. *Sensors and Actuators B: Chemical*, *202*, 426-432.

Callis, J. B., Illman, D. L., & Kowalski, B. R. (1987). Process analytical chemistry. *Analytical Chemistry*, *59*(9), 624A-637A.

Engelsen, S. B. (1997). Explorative spectrometric evaluations of frying oil deterioration. *Journal of the American Oil Chemists' Society*, *74*(12), 1495.

Zachariassen CB, Larsen J, van den Berg F, Engelsen SB. Use of NIR spectroscopy and chemometrics for on-line process monitoring of ammonia in low methoxylated amidated pectin production. Chemometr. Intell. Lab. Syst. 2005; 76: 149–161.

Salomonsen, T., Jensen, H. M., Stenbæk, D., & Engelsen, S. B. (2008). Chemometric prediction of alginate monomer composition: A comparative spectroscopic study using IR, Raman, NIR and NMR. *Carbohydrate Polymers*, *72*(4), 730-739.

Sørensen, K. M., Petersen, H., & Engelsen, S. B. (2012). An on-line near-infrared (NIR) transmission method for determining depth profiles of fatty acid composition and iodine value in porcine adipose fat tissue. *Applied spectroscopy*, *66*(2), 218-226.

Vo, A.Q., He, H., Zhang, J., Martin, S., Chen, R., Repka, M.A., 2018. Application of FT-NIR analysis for in-line and real-time monitoring of pharmaceutical hot melt extrusion: a technical note. Aaps Pharmscitech. https://doi.org/2010.1208/s12249-12018-11091-12243.

Wesholowski, J.; Prill, S.; Berghaus, A.; Thommes, M. Inline UV/Vis Spectroscopy as PAT tool for Hot-Melt Extrusion. Drug Deliv. Transl. Res. **2018**, 1, 1–9.

Crouter, A. & Briens, L. Methods to Assess Mixing of Pharmaceutical Powders. AAPS PharmSciTech (2019) 20: 84. https://doi.org/10.1208/s12249-018-1286-7.

Pauli, V., Roggo, Y., Pellegatti, L., Trung, N. Q. N., Elbaz, F., Ensslin, S., ... & Krumme, M. (2019). Process analytical technology for continuous manufacturing tableting processing: A case study. *Journal of pharmaceutical and biomedical analysis*, *162*, 101-111.

Trenfield, S. J., Goyanes, A., Telford, R., Wilsdon, D., Rowland, M., Gaisford, S., & Basit, A. W. (2018). 3D printed drug products: Non-destructive dose verification using a rapid point-and-shoot approach. *International journal of pharmaceutics*, *549*(1-2), 283-292.

Chapter 5

Case studies research

5.1 Pharmaceutical industry

5.1.1 Introduction

The manufacturing process of dietary supplements and food can be outlined in the following consecutive stages: raw materials receiving, inspection of the incoming materials, production, packaging and the final product delivery. The conformity of incoming materials can be assessed through sampling and the following analysis according to established and planned procedures based on the good manufacturing practices (GMPs). One usable way to assess the quality of materials involves the use of conventional, reference methods established by regulatory bodies. These analytical methods besides of being reliable have some disadvantages such as the need of sample treatment, long time of analysis and the related material and environmental costs. Moreover, the time consuming reference methods cannot meet the timely outcomes expected by the suppliers. The delay in the acceptance procedure has adverse effects for those raw materials that has to be stored under proper conditions and disadvantages for *just in time* manufacturing procedures. The delay in the analytical response can be reduced by the implementation of rapid, non-destructive testing methods like NIR spectroscopy in combination with computational methods. The main purposes of the conformity control of incoming raw materials consist in preventing the use of out-of-specification materials, identification of possible adulteration and the assurance that the materials fulfil the recipe requirements (Hitzmann, 2015). The high frequency quality control of incoming materials and ingredients as an integrated part of the PAT can be performed by NIR technology according to an untargeted approach in order also to detect deliberate adulteration. The above mentioned potential of NIRS is just an integration of the capability to perform targeted calibrations of quality attributes (Sorensen, 2016). The non-conformity of starting materials due to several economic and/or effectual adulterants such as Melamine testing which was recommended by FDA led the dietary supplement manufacturers to the development of testing procedures for all the incoming raw materials as they are responsible for the quality of their released products (Champagne, 2011).

Regarding the botanical raw materials (botanicals), the presence of intentional and unintentional adulterants like active substances or other compounds arising from the unwanted part of the plant or from a different plant can result in toxic

effects or cannot provide the expected pharmacological effects. Moreover, a large number of naturally occurring contaminants in raw materials can be a source of non-conformity with adverse effects on the intended use of the end product (Sanzini, 2011).

NIRS and chemometrics proved to be useful in the discrimination of several celluloses with different physical and chemical properties, employed as excipients for pharmaceutical purposes, providing reliable identification (Krämer, 2000).

Raw material testing based on NIRS demonstrated its potential benefits not only in pharmaceutical organizations preventing unexpected products. A raw material spectral library built with spectroscopic data can be effectively used for identity confirmation and classification purposes taking advantage of the created models based on qualitative algorithms (Kemper, 2003).

A proactive control approach of the process is accomplished when the output of testing of the raw materials is used with the aim of determining the process conditions to assure the product quality. The fulfilment of product specifications and the reduction or elimination of waste (muda) can be attained by continuous process monitoring and control which permits corrective actions during the process instead of the reactive approach where the information on product quality is used to undertake the corrective actions on the process (Kourti, 2006).

The purpose of the study concerns: 1. The development of chemometric models for identity confirmation and identification (classification) which allow to compare the NIR spectrum of a material (known or unknown) with the spectra of known materials employed previously for qualitative modelling; 2. The development of a regression model based on the PLS method in order to predict the quantity of the active ingredient DHA (docosahexaenoic acid) in a semi-finished or end solid product.

5.1.2 Materials and methods

5.1.2.1 Raw materials and semi-finished product

Raw materials. The materials and ingredients under investigation have been employed as active compounds and excipients in the formulation and manufacturing of end products (dietary supplements) and has been supplied from several manufacturers over a period of three years. A list of the examined solid materials including botanicals, pure substances and other solids is displayed as follows (Table 5-1). Moreover, the NIR spectra of other raw materials have been collected but not subjected to modelling due to the limited number of lots for each of these materials.

Table 5-1. List of the botanicals and other solid raw materials with the highest number of lots purchased by the company (*Raw materials subjected to exploratory analysis and classification purposes that have been included in this work).

Botanicals (dry extracts)			
	Name	Lots	Note
1	Acerola	8	
2	Alga wakame	6	
3	Aloe	8	
4	Bergamot	17	*
5	Hawthorn	7	
6	Boswellia	12	*
7	Eschscholzia	12	
8	Fennel	10	
9	Fucus vesiculosus	17	
10	Ginkgo biloba phytosoma	8	*
11	Ginkgo biloba 6%	4	*
12	Ginkgo biloba 24%	17	*
13	Ginseng	10	
14	Melilotus	10	*
15	Melissa 0.5%	7	*
16	Melissa 2%	10	*
17	Melissa 6%	6	*
18	Passionflower 1-1.2%	16	
19	Passionflower 4%	9	
20	Passionflower parti aeree	7	
21	Propolis	7	
22	Soybean	20	*
23	Green tea	7	
24	Tilia	9	
25	Valerian	13	*

26	Ginger	9	
27	Senna tinnevelly	17	
Pure substances and other solids			
28	Acesulfame K	7	
29	Acetyl-L-carnitine hydrochloride	18	
30	Anhydrous citric acid	14	*
31	Monohydrate citric acid	6	*
32	Hyaluronic acid	13	
33	Thioctic acid Matris fast	25	*
34	Thioctic acid Matris retard	25	*
35	Alpha galactosidase	16	*
36	Beta galactosidase	14	*
37	Orange aroma CC	10	*
38	Orange aroma spray	10	*
39	Lemon drycell aroma	11	*
40	Lemon juice aroma	9	*
41	Passion fruit aroma	11	
42	Bamboo fibre	9	
43	Beta carotene powder	11	
44	Anhydrous caffeine	10	
45	Cetyl alcohol	9	
46	Citicolin	20	
47	Choline bitartrate	13	
48	Creatine monohydrate	10	
49	Cutin	10	
50	Corn anhydrous dextrose	16	
51	Docosahexaenoic acid	14	
52	Diosmin	16	

53	Hesperidin	16	
54	Konjac flour glucomanan	20	
55	Fe gluconate	9	
56	Fructooligosacharides	10	
57	Fructose	21	
58	Kollidon	9	
59	Lactium	15	
60	Lactoferrin	14	
61	L-arginine base	13	*
62	L-arginine hydrochloride	16	*
63	Whole milk in powder	10	
64	Lactose	10	
65	Lecithin	10	
66	L-glutamine	15	
67	Libramed	57	*
68	Mg hydroxide	11	
69	Mg oxide	24	
70	Mg pidolate	11	
71	Maltodextrin	20	
72	Dried maltodextrin	15	
73	Mannitol P 400	11	*
74	Mannitol P SD 200	15	*
75	Melatonin	11	
76	Vitaminic mix	12	
77	K citrate	14	*
78	K sorbate	12	*
79	Na benzoate	9	
80	Fermented soybean	19	
81	Sorbitol	28	
82	Stearyl alcohol	11	

83	Troxerutin	21	
84	Ubidecarenone Q10	12	*
85	Ubidecarenone 10%	50	*
86	Vitamin A acetate	16	
87	Vitamin B12 0.1%	12	
88	Vitamin B5	9	
89	Vitamin B6	9	
90	Vitamin C	38	
91	Vitamin E 50%	11	
92	Xangold beadlets 10%	17	*
93	Xangold beadlets 20%	13	*
94	Xilitol xilisorb	22	
95	Zeaxanthin	15	

The samples withdrawn from the batches of purchased raw materials have been put within opaque glass containers, stored at controlled temperature (25°C) and humidity (30%) conditions before analysis.

Semi-finished product. Concerning the potential of NIRS in the quantification of DHA in the semi-finished product, 22 mixtures including a replicate have been prepared taking into account the mixing process carried out by the company and the specification limits. The total number of concentration levels in DHA has been 11 ranging from 5% to 10%. The partial list of the solid ingredients used for the product formulation is shown in the following table (Table 5-2). The starting raw material is a complex mixture containing DHA.

Table 5-2. Incomplete list of the solid raw materials employed to make the product.

DHA
Citric acid
X4
Sucralose
Zinc citrate
X5
X2
X1
Isomalt Galeniq
X3
Sorbitol

5.1.2.2 Data collection

The spectral data of both the incoming raw materials and the mixtures with different concentrations in DHA were acquired using a FT-NIR spectrometer (Buchi NIRFlex N-500) equipped with a fiber optic probe for solids (Figure 5-1). The following acquisition parameters were employed:

Acquisition modality: Diffuse reflectance

Spectral range: 10000 – 4000 cm^{-1}

Resolution: 8 cm^{-1}

Detector: InGaAs photodiodes

Figure 5-1. Buchi FT-NIR instrument used to collect the spectral data from the abovementioned samples.

Three spectra have been collected for each sample, in three different spots, by placing the probe in contact with the solid material. Before sample analysis the background has been acquired by reflecting the entire range of selected radiation.

5.1.2.3 Multivariate data analysis

The raw spectral data once acquired have been subjected to several pre-processing techniques and their combinations such as mean centring, SNV, MSC, Savitzky–Golay algorithm, etc. in order to reduce or eliminate the variations not related to the chemical information of spectra. Afterwards, the pre-processed spectral data have been modelled by using pattern recognition methods like PCA algorithms and classification methods such as PLS-DA algorithms with the aim of incoming raw materials conformity assurance. Likewise, the spectra arising from the solid samples with different concentrations in DHA have been pre-processed and subjected to regression methods like PLS. To achieve the abovementioned

objectives have been employed the LatentiX software and PLS toolbox used within the MATLAB environment.

5.1.3 Results and discussion

Raw materials. The qualitative models developed, in order to monitor the conformity of the incoming raw materials, can be implemented by the company in order to have a rapid conformity assessment allowed by the NIRS. This approach i.e. the use of NIRS and chemometrics allows the organization to assess the quality of the materials provided by the suppliers, even if they are accompanied with the certificate of analysis, as the main responsibility for the non-compliant end products pertains the manufacturing company. On the other hand, the use of time consuming analytical techniques will result in a production delay. The methods used showed the potential to build models, which allow identity confirmation or classification, either for pure substances or for complex ingredients such as dry extracts which consist of a mixture of compounds including additives as well.

Once the model has been developed, the limits of a known set of raw materials used for identity confirmation can be assigned by calculating the Mahalanobis distance allowing in this way the recognition of doubtful samples and outliers. The Figure 5-2 displays the PCA model of Libramed (raw material) on which the limits of this group of samples are provided according to the Mahalanobis distance. The plot shows that there are some doubtful amber colored samples which most likely belong to Libramed besides the presence of large outliers colored in purple.

This concept can be extended to pattern recognition or classification models depicting more than one different group or class.

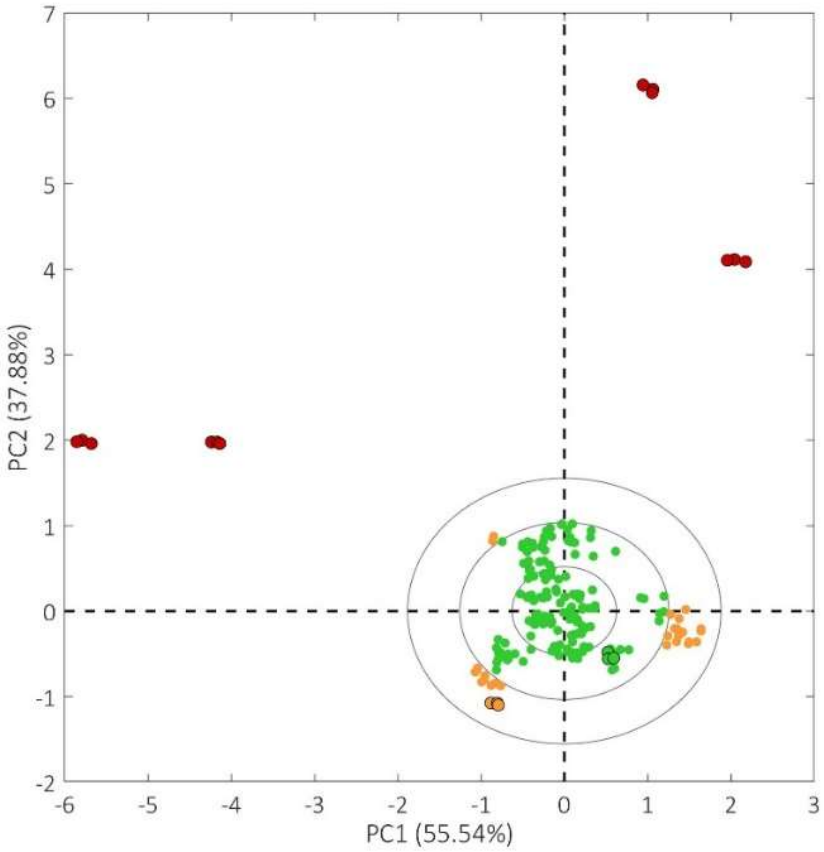

Figure 5-2. PCA scores plot depicting the limits of the known group of libramed samples based on the Mahalanobis distance.

The company up to now succeeded in the development of a few models related to the pure incoming raw materials whereas the models developed during this study covered also the more complex ingredients.

The following figures display only the PCA models, stemming from the NIR spectral data considering the whole spectral range (between 10000-4000 cm-1), as the classification models developed using the PLS-DA method provide a similar result.

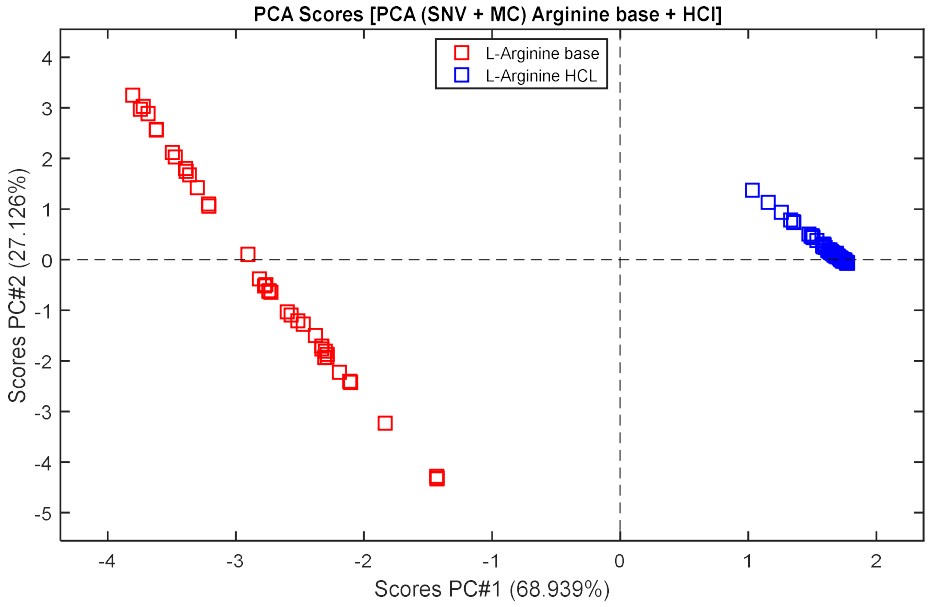

Figure 5-3. PCA scores plot of arginine base and hydrochloride using SNV and MC as pre-processing techniques.

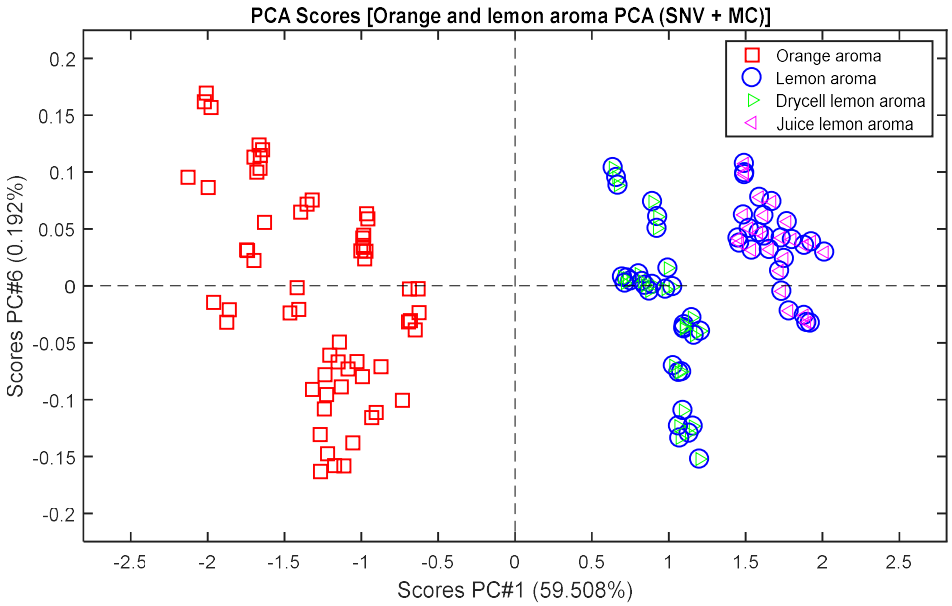

Figure 5-4. PCA scores plot of orange and lemon aroma using SNV and MC as pre-processing techniques.

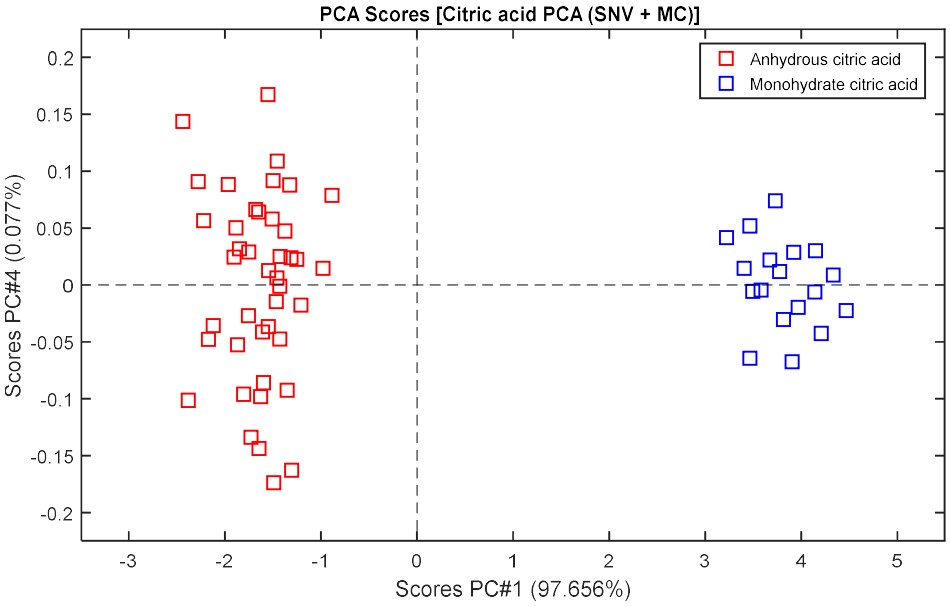

Figure 5-5. PCA scores plot of anhydrous and monohydrate citric acid using SNV and MC as pre-processing techniques.

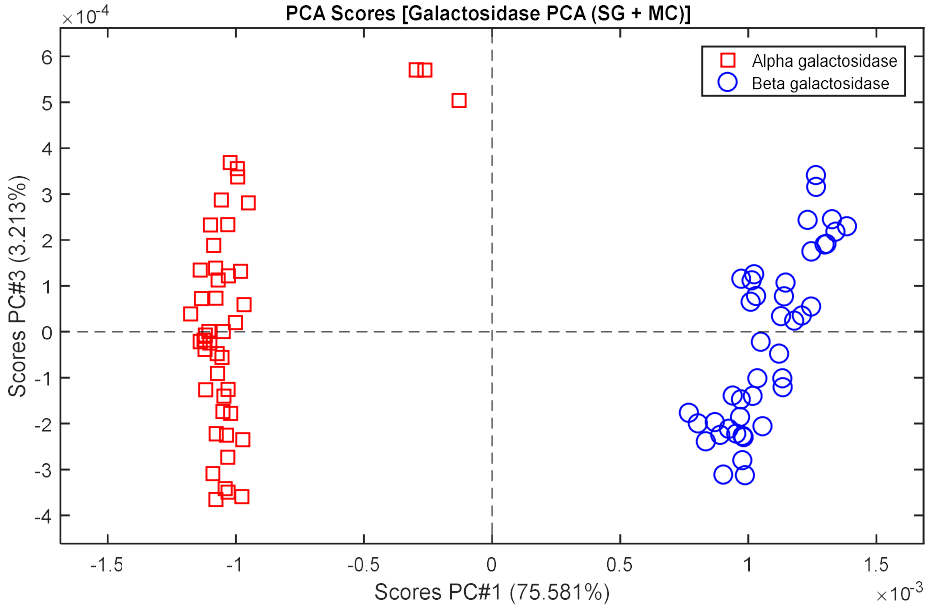

Figure 5-6. PCA scores plot of alpha and beta galactosidase using Savizky-Golay (window size: 19; polynomial order: 2; 2nd derivative) and MC as pre-processing techniques.

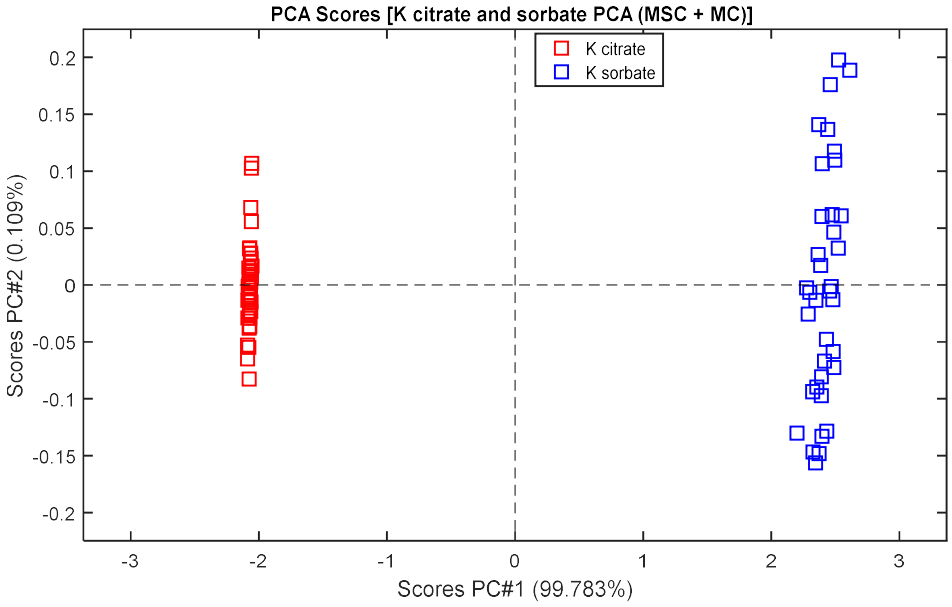

Figure 5-7. PCA scores plot of potassium citrate and sorbate using MSC and MC as pre-processing techniques.

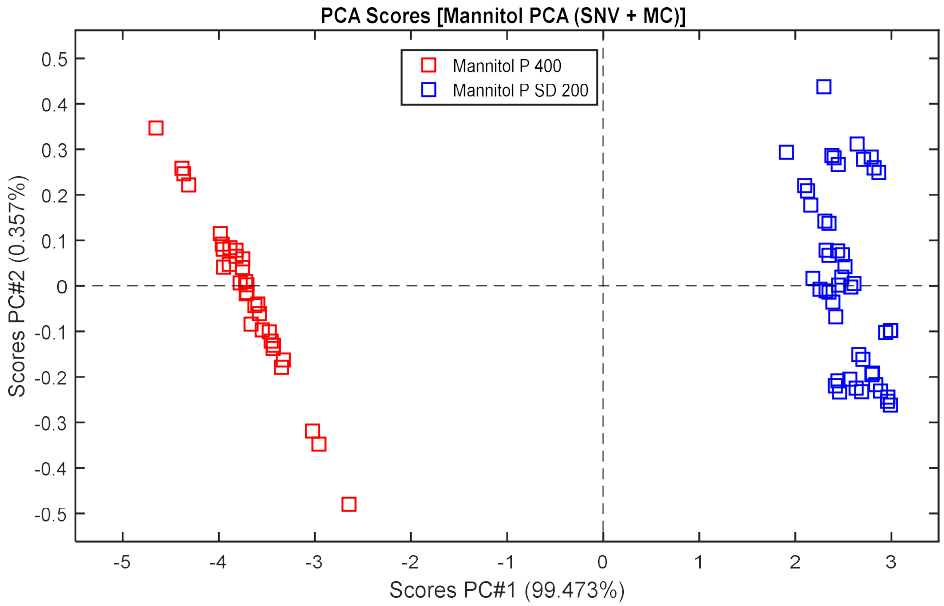

Figure 5-8. PCA scores plot of two grades mannitol, according to particles size, using SNV and MC as pre-processing techniques.

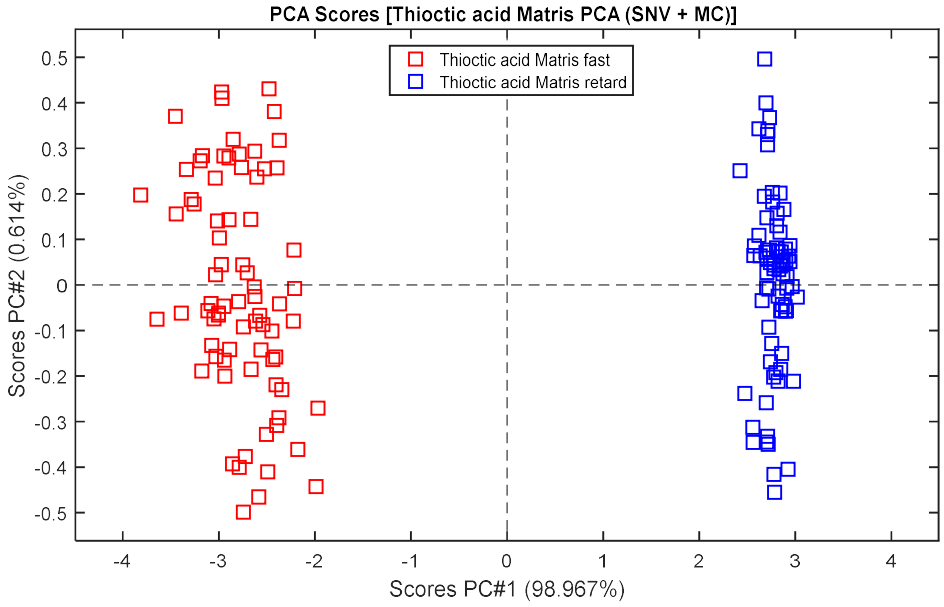

Figure 5-9. PCA scores plot of thioctic acid in the two forms (Matris fast and retard) using SNV and MC as pre-processing techniques.

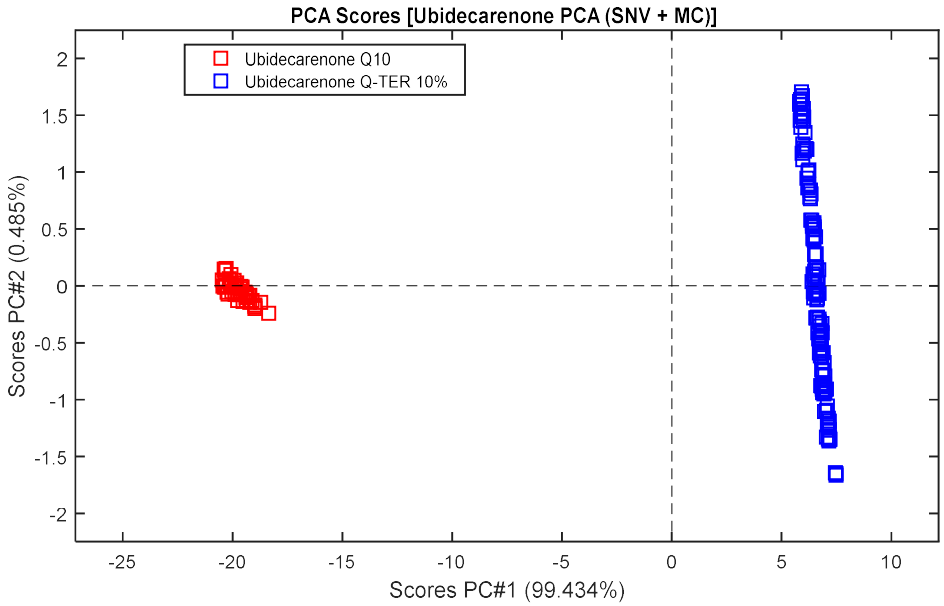

Figure 5-10. PCA scores plot of ubidecarenone in the two forms using SNV and MC as pre-processing techniques.

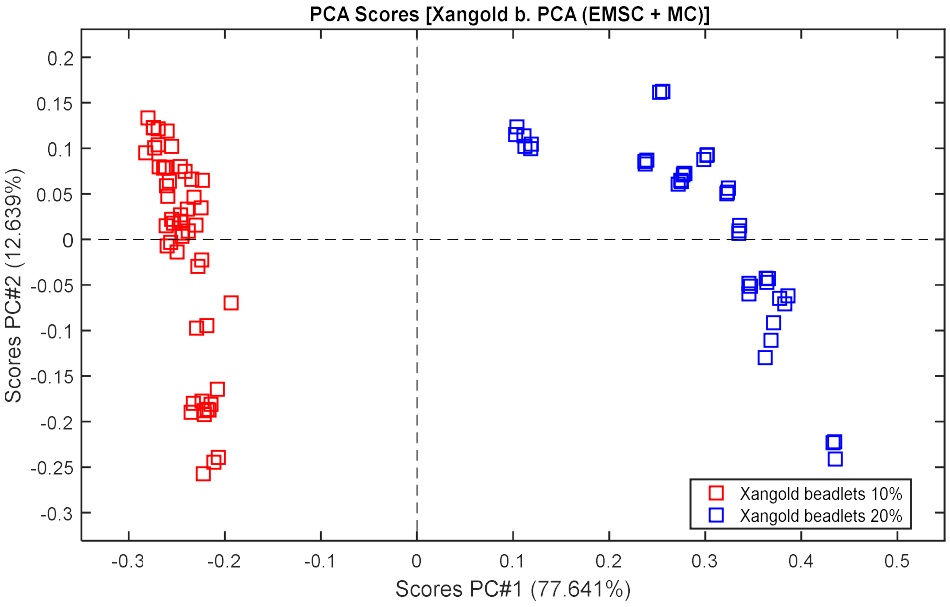

Figure 5-11. PCA scores plot of xangold in the two forms using EMSC and MC as pre-processing techniques.

Among the solid materials, the identity confirmation or classification of botanicals were of particular interest for the company as they are more complex and belong mostly to the active ingredients in the dietary supplement recipes. Some PCA models are displayed in the following figures.

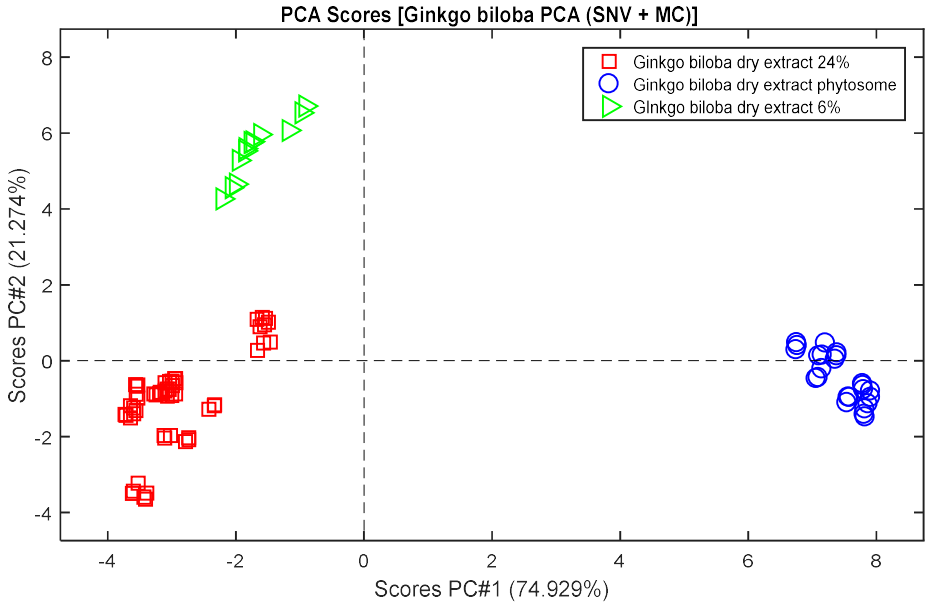

Figure 5-12. PCA scores plot of the three forms of ginkgo biloba dry extracts using SNV and MC as pre-processing techniques.

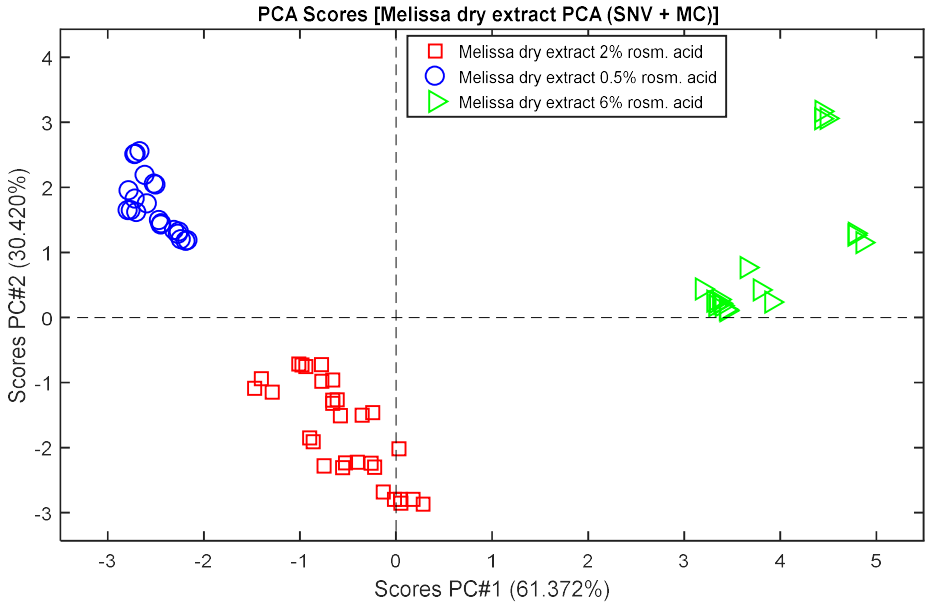

Figure 5-13. PCA scores plot of the three forms of melissa dry extracts using SNV and MC as pre-processing techniques.

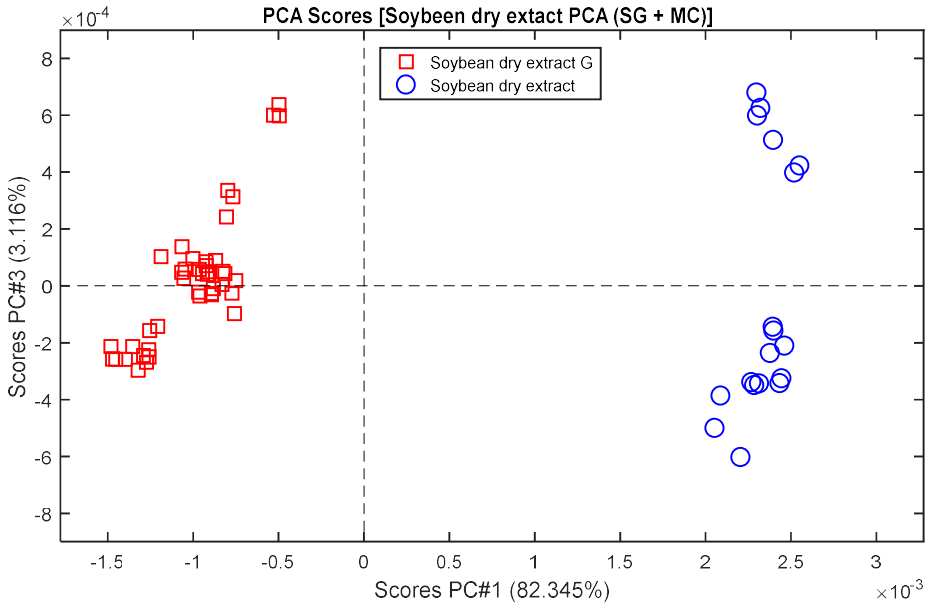

Figure 5-14. PCA scores plot of soybean dry extracts, purchased from two suppliers, using SG (window size: 19; polynomial order: 2; 2nd derivative) and MC as pre-processing techniques.

The scores plot displayed above shows predominantly two groups of soybean dry extracts with the same denomination which are well separated according to the two different suppliers.

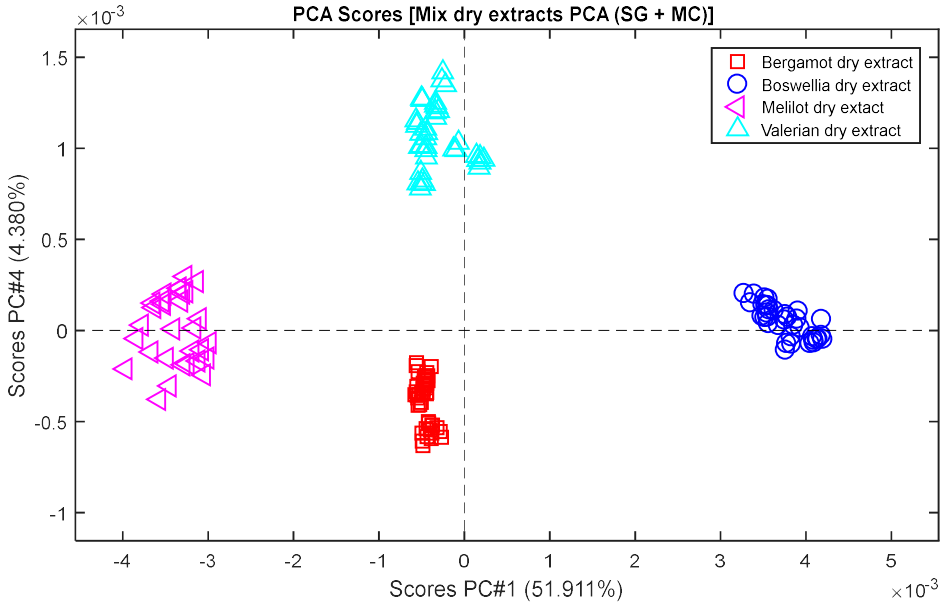

Figure 5-15. PCA scores plot of four different dry extracts using SG (window size: 19; polynomial order: 2; 2nd derivative) and MC as pre-processing techniques.

An example of model implementation performed by the staff of the organization (aizoon Technology Consulting) concerns the Ginkgo biloba dry extract in three different chemical compositions mostly related to the concentration of active constituent/s. This qualitative model resulted from classification methods (PLS-DA) has been first developed starting from the spectral data collected and then implemented in a dashboard through a Data Lake. The NIR spectral data, belonging to new Ginkgo biloba 24% dry extracts (not included in the training set), collected in few seconds have been synchronized and uploaded in the dashboard within which have been automatically pre-processed and predicted through the classification model displaying successfully the membership class as shown in Figure 5-16.

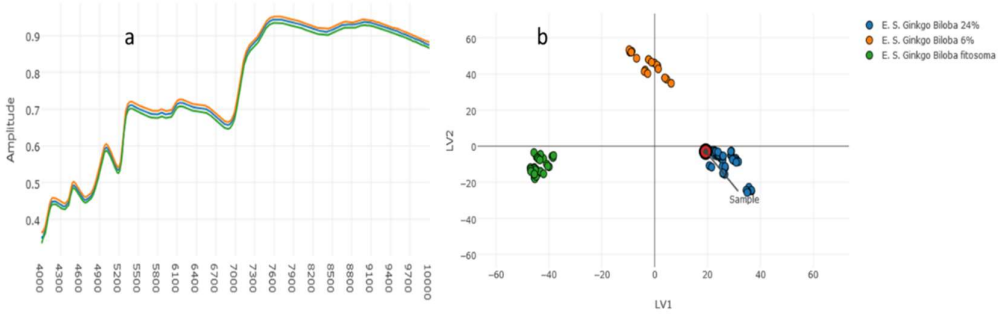

Figure 5-16. Ginkgo biloba 24% pre-processed spectra (a) and their subsequent prediction on the previously developed PLS-DA model (b).

Semi-finished product. NIRS coupled with chemometrics displayed also the potential related to the quantitative prediction of the DHA as active constituent in the semi-finished and end product. To accomplish this goal, the PLS regression algorithm has been applied to the pre-processed (SNV + MC) spectral data and the model obtained is shown in Figure 5-17. The regression model chosen with 4 principal components, as the RMSEs of cross validation (0.31) using Venetian Blinds and calibration (0.30) were close, resulted in a R^2 for cross validation of 0.96.

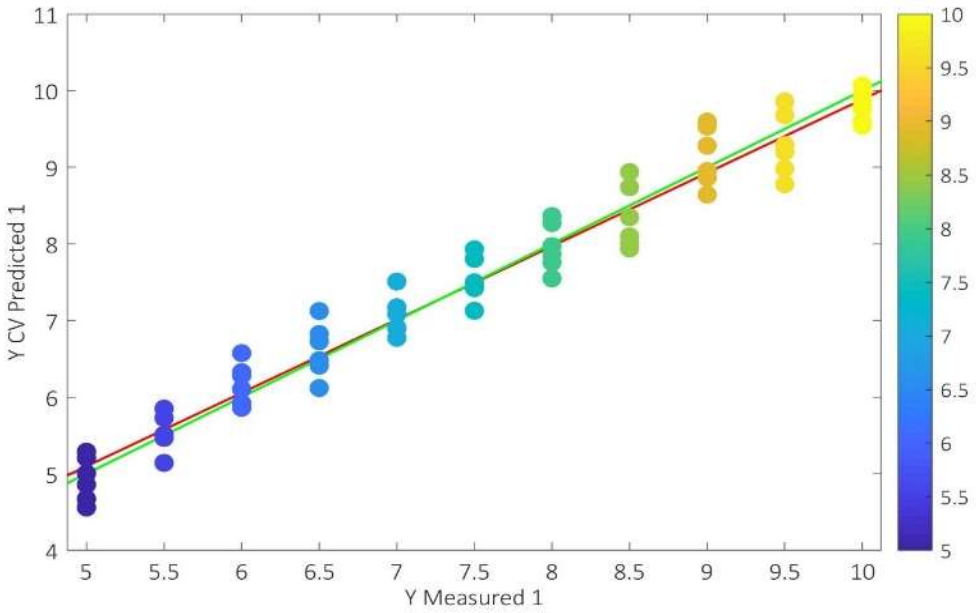

Figure 5-17. PLS linear regression model depicting the predicted versus the actual concentration of DHA in the various mixtures ranging from 5% to 10% according to the DoE explained in Section 5.1.2.

The quantity of DHA in two end products manufactured in different times has been predicted, using this model in which the mean spectrum of the three collected spectra of each standard mixture has been considered, showing the allocation of samples (marked in red) within the specification limits, i.e. 6.6% and 9% (Figure 5-18).

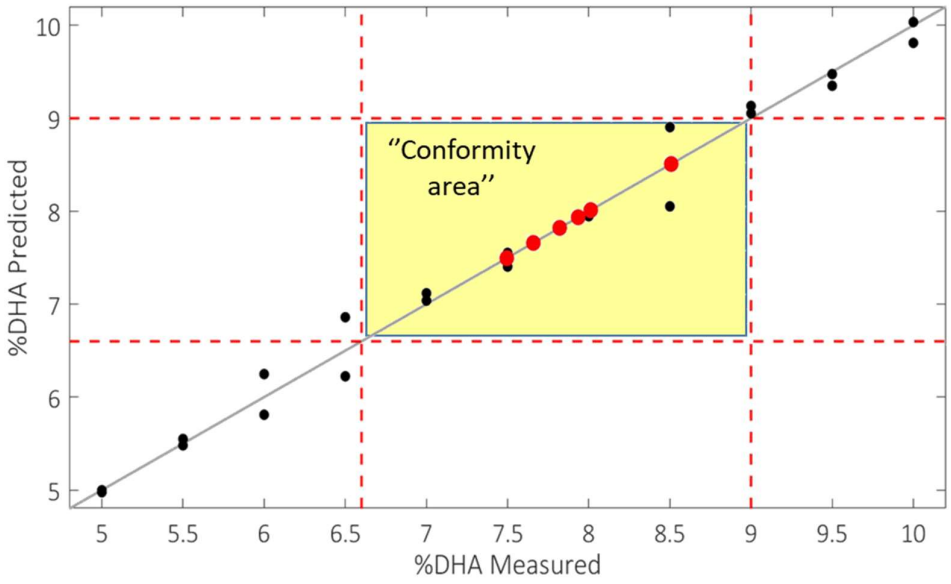

Figure 5-18. Prediction of six NIR spectra representing two final products manufactured in two different times.

It has also predicted the concentration of DHA in two prepared out of specification samples having a higher and lower title with respect to the specification limits. Through the linear regression model these samples have been predicted out of the specification limits.

The NIR spectra of another end product without DHA have been collected and after pre-processing have been predicted using this model. This end product has been placed on the regression model far away from the specification limits.

5.1.4 References

Hitzmann, B., Hauselmann, R., Niemoeller, A., Sangi, D., Traenkle, J. & Glassey, J. (2015). Process analytical technologies in food industry - challenges and benefits: a status report and recommendations. Biotechnology Journal, 10, 1095–1100.

Sørensen, K. M., Khakimov, B., & Engelsen, S. B. (2016). The use of rapid spectroscopic screening methods to detect adulteration of food raw materials and ingredients. *Current Opinion in Food Science*, *10*, 45-51.

Champagne, A. B., & Emmel, K. V. (2011). Rapid screening test for adulteration in raw materials of dietary supplements. *Vibrational Spectroscopy*, *55*(2), 216-223.

Sanzini, E., Badea, M., Dos Santos, A., Restani, P., & Sievers, H. (2011). Quality control of plant food supplements. *Food & function*, *2*(12), 740-746.

Krämer, K., & Ebel, S. (2000). Application of NIR reflectance spectroscopy for the identification of pharmaceutical excipients. *Analytica chimica acta*, *420*(2), 155-161.

Kemper, M. S., & Luchetta, L. M. (2003). A guide to raw material analysis using near infrared spectroscopy. *Journal of near infrared spectroscopy*, *11*(3), 155-174.

Kourti, T. (2006). Process analytical technology beyond real-time analyzers: The role of multivariate analysis. *Critical reviews in analytical chemistry*, *36*(3-4), 257-278.

5.2 Vegetable oil company

5.2.1 Introduction

The continuous study on vegetable oils related to their ever increasing employment for various purposes in food, pharmaceutical and other sectors (Spatari, 2017) is due to their benefits in providing essential fatty acids, liposoluble vitamins and other quality attributes (Li, 2016).

The quality of edible oils can be determined based on physicochemical parameters such as free acidity, peroxide value, anisidine value. Acidity value, expressed as the percentage in grams of oleic acid, is a very important parameter mostly for olive oils as their classification is based on free fatty acid content according to the European Regulation and the product intended for marketing cannot be subjected to neutralization. Peroxide value, determining the amount of hydroperoxides resulted by a radical mechanism of degradation over the oil exposure to light, heating, air, etc., is an indicator of the primary oxidation of edible oils expressed as the milliequivalents of peroxide in a kilogram of oil. Anisidine value, instead, measures the secondary oxidation arising from the decomposition of peroxides which mainly results in carbonyl compounds. Edible oils and fats consist on acylglycerols (esters of glycerol with fatty acids) where the triglycerides represent the most abundant structures. Fatty acids, according to the presence or not of the double bonds, are distinguished in saturated, monounsaturated and polyunsaturated, can have a specific role on health depending on their quality and quantity (Dorni, 2018).

Besides the traditional methods of analysis based mainly on titration, several laboratory studies showed the potential in quantifying free acidity and peroxide index in olive oils and other edible oils by means of NIRS (Armenta, 2007). The classification of several edible oils as well as quantitative information related to fatty acids profile have been performed by a developed rapid NIR method (Azizian, 2005). The classification of extra virgin olive oils based on their geographical origin has been carried out by means of a UV-VIS-NIR spectroscopy equipped with a fibre optic (Mignani, 2007). Authentication and adulteration studies on edible oils have been performed using NIRS and chemometrics (Armenta, 2010).

Additionally, other parameters used to assess the quality of edible oils include sensory characteristics like taste, aroma and colour. The oil colour which relies on the quantity and category of pigments is one of the characteristics which affects consumer's acceptance. Pigments such as chlorophylls and carotenoids, which occur in oilseeds and accordingly in the crude oil, are mostly responsible for the colour. The amount of these pigments is determined by the type of raw vegetable oil and the processing techniques used for the production (El-Hamidi, 2016). Light and heat are the main sources of pigments deterioration, particularly chlorophylls, modifying the sensory characteristics (Choe, 2006).

Several studies have been carried out in the pharmaceutical sector, focused on oil photo-stability due to the manifested limits in pharmacological activity and the evolution in toxic products. Food products containing oil can even reduce or miss their sensory properties and nutrients during the contact with light (Spatari, 2017).

The quantification of the abovementioned quality and sensory parameters determines the freshness and the shelf life of several types of oils.

The purposes of the present study on vegetable oils concern: 1. The assessment of the shelf life of vegetable oils; 2. Quality evaluation of extra virgin olive oils as a function of the storage time; 3. The effect of two cold pressing systems on five vegetable oils produced starting from their seeds; 4. The evaluation of cold-pressed linseed oil oxidative stability when subjected to accelerated oxidation.

5.2.2 Materials and methods

5.2.2.1 Materials and designs

Shelf-life assessment. Three different types of vegetable oils have been employed in order to assess their shelf life. Hemp oil, linseed oil and the sunflower oil have been yielded via mechanical extraction using an expeller press (cold pressing equipment) starting from their seeds. The oils have been extracted within the same day using the same pressing conditions in terms of press speed and put in dark glass bottles having a volume of 250 millilitres. In order to assess their shelf-life the following experimental design have been made: the bottles stored at 20°C have been subjected to artificial radiation for a period of 12 months, employing a NEON (Cool White 840) and a LED light with the same light power and colour (6500K and 1500 lumen), switching between the light (12 hours) and the dark (12 hours). At time intervals of two months one bottle for each sample have been withdrawn from the storage and subjected to several analyses using primary and secondary analytical techniques.

Extra virgin olive oil. Regarding the extra virgin olive oil, two batches have been stored in two tanks. The first tank containing an already stored oil by 6 months and the second one containing a fresh oil (just stored). From each tank, at intervals of 15 days, two samples have been withdrawn (one from the upper part and the other one from the lower part of the tank) noticing the external and internal tank temperature and afterwards have been subdued to testing in order to assess the oil quality.

Comparison between two cold pressing systems. Concerning the third point, an expeller press and a hydraulic piston press have been used for pressing the several seeds (hemp, linseed, sunflower, pumpkin and walnut). The expeller press operated under two pressing conditions i.e. cochlea speed at 70% and at 100%. Whereas the two working conditions of the hydraulic press relied on the applied pressure (860 kg/cm^2 and 900 kg/cm^2) and on the pressing time (700" and 900") as displayed in Table 5-3. Unlike walnut where two pressing replicates have been accomplished for each condition, the remaining oils have been yielded in three replications for each condition.

Table 5-3. Working conditions of the mechanical expeller pressing and hydraulic press.

	Expeller press	Hydraulic piston press
Condition 1	70%	860 kg/cm^2 – 700"
Condition 2	100%	900 kg/cm^2 – 900"

Cold-pressed linseed oil oxidative stability. With respect to the last objective, an expeller press operating at two different speeds (20% and 80%) has been employed for pressing a linseed batch. For each of the two operating conditions have been performed three pressing replicates. During each pressing an assembled portable FT-NIR tool has been employed to collect the NIR spectra in real time as shown in Figure 5-19. The linseed oil, after pressing, has been subjected to accelerated oxidation for five days in order to assess its oxidative stability. The NIR spectra of linseed oil before and after oxidation have been acquired using the bench FT-NIR instrument. The same modality of analysis has been used for the other abovementioned oils as well.

Figure 5-19. Expeller press equipped with a thermocouple and an in-house assembled portable FT-NIR instrument equipped with a fibre optic transflectance probe.

5.2.2.2 Data collection

The spectral data have been acquired without sample pre-treatments using two spectroscopic techniques: 1. FT-NIR MPA spectrometer (Bruker optics, Ettlingen, Germany) equipped with a sample compartment for liquid samples testing (Figure 5-20a). The samples have been placed in a transparent glass vial before the acquisition of spectra; 2. UV-VIS spectrometer (Cary 5000, Agilent) displayed in Figure 5-20b). The cuvettes in polystyrene have been used as sample holder.

The acquisition parameters employed for spectra collection by FT-NIR are displayed as follows:

Acquisition modality: Absorbance
Spectral range: 12500 - 4000 cm^{-1}
Resolution: 16 cm^{-1}
Detector: InGaAs photodiodes
Sample scans: 35
Optical path: 8 mm
Acquisition temperature: 35°C

Figure 5-20. FT-NIR MPA spectrometer from Bruker optics (a) and Varian Cary 5000 UV-Vis Spectrophotometer (b).

The characteristics of UV-VIS method used to acquire the spectral data are listed below:

UV-VIS source: Deuterium-Tungsten halogen

Acquisition modality: Absorbance

Spectral range: 300 - 800 nm

Optical path: 1 cm

Detector: Photomultiplier tube

An example of NIR spectrum of extra virgin olive oil is displayed in Figure 5-21.

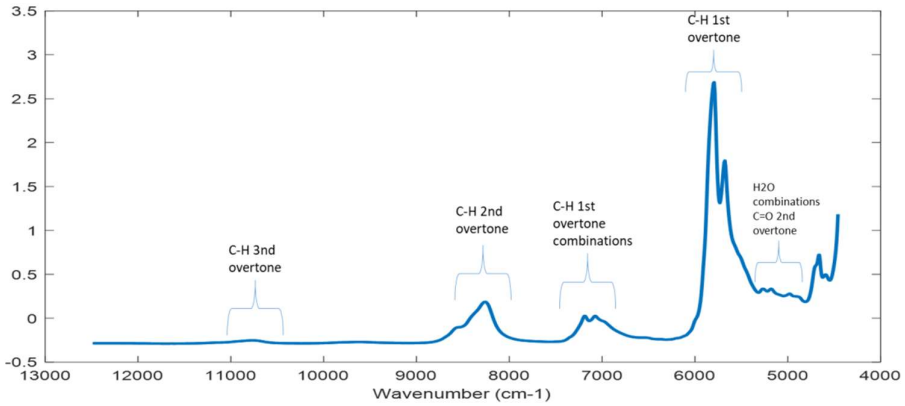

Figure 5-21. NIR spectrum of olive oil with the assignment of the main absorption bands.

5.2.2.3 Multivariate data analysis

The raw spectral data collected using NIRS and UV-VIS spectroscopy have been subjected to pre-processing employing in combination SNV + MC techniques. Afterwards, the pre-processed spectral data have been analysed by using

exploratory data analysis (PCA algorithm) employing the PLS toolbox within the MATLAB environment.

5.2.3 Results and discussion

Shelf-life assessment. The three different vegetable oils employed for their quality and shelf life assessment showed different behaviour along the storage in controlled conditions. The colour of the linseed and sunflower oils had not changed in the first six months of light exposure. On the contrary concerning the hemp oil after two months of light exposure the effect of the two light sources on the colour was minimal compared to the initial colour (green colour due to chlorophyll pigments with absorbances at 664 nm (chlorophyll *a*), 647 nm (chlorophyll *b*) and 631 nm (chlorophyll c_1 and c_2) taking into account the purified pigments (Lorenzen, 1980)). Moreover, there was also a little difference in the colour due to the two light sources. After four months of NEON light exposure the colour of hemp oil shifted towards yellow due to carotenoid pigments with absorbances at 432, 455 and 480 nm (El-Hamidi, 2016) whereas the effect of LED apparatus was less apparent. The NEON light showed a more drastic effect in chlorophylls deterioration after four months compare to LED. After six months of NEON light treatment the colour of hemp oil became almost yellow and the impact of LED yielded a yellow colour effecting in this way the chlorophylls transformation (Figure 5-22).

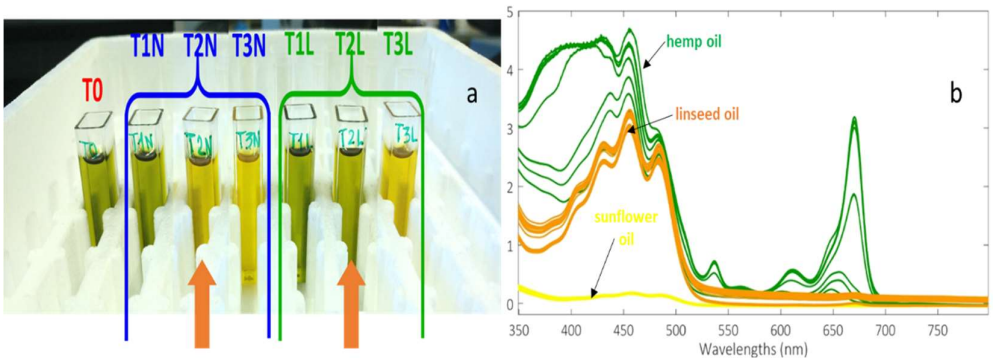

Figure 5-22. Visible impact of hemp oil arising from NEON and LED light exposure (a) and the spectra of the three oils collected in the visible region (b).

The visible impact of pigment transformation can be explained by modelling the spectral data arising from both NIR analysis (Figure 5-23) and UV-VIS spectroscopy (Figure 5-24) using exploratory data analysis. From the PCA model, developed from the spectral data collected in the visible region, it is clear that the hemp oil after four months under NEON light is close to the samples exposed for a period of six months on both the lights whereas the scores of the same hemp oil exposed to LED are close to the scores of the fresh hemp oil.

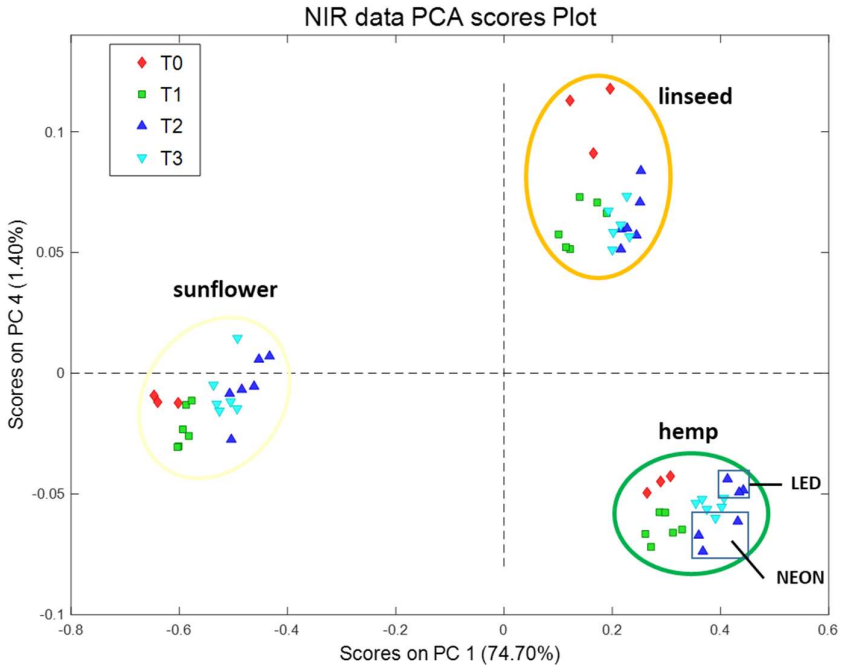

Figure 5-23. PCA scores plot of NIR spectral data of the three vegetable oils.

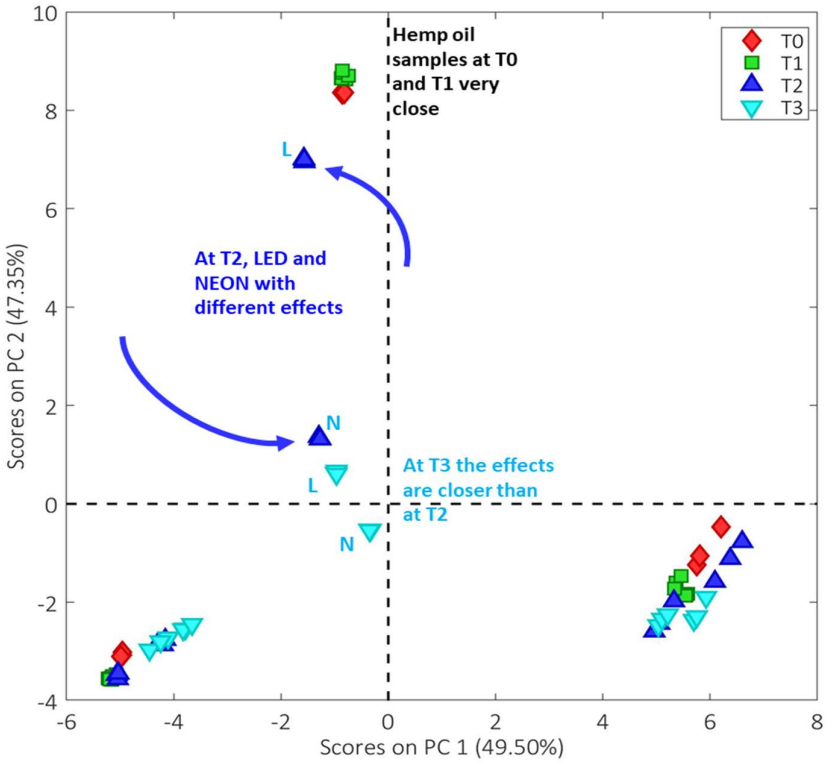

Figure 5-24. PCA scores plot of visible spectral data of the three vegetable oils over the six months of storage under controlled conditions.

Extra virgin olive oil. The extra virgin olive oils withdrawn from the two tanks (silos) over a period of five months (from August to December) and analysed by NIRS turn out to be distinguished by the scores plot of the PCA model (Figure 5-25). This difference may arise as a result of the two different cultivars or due to the different storage time.

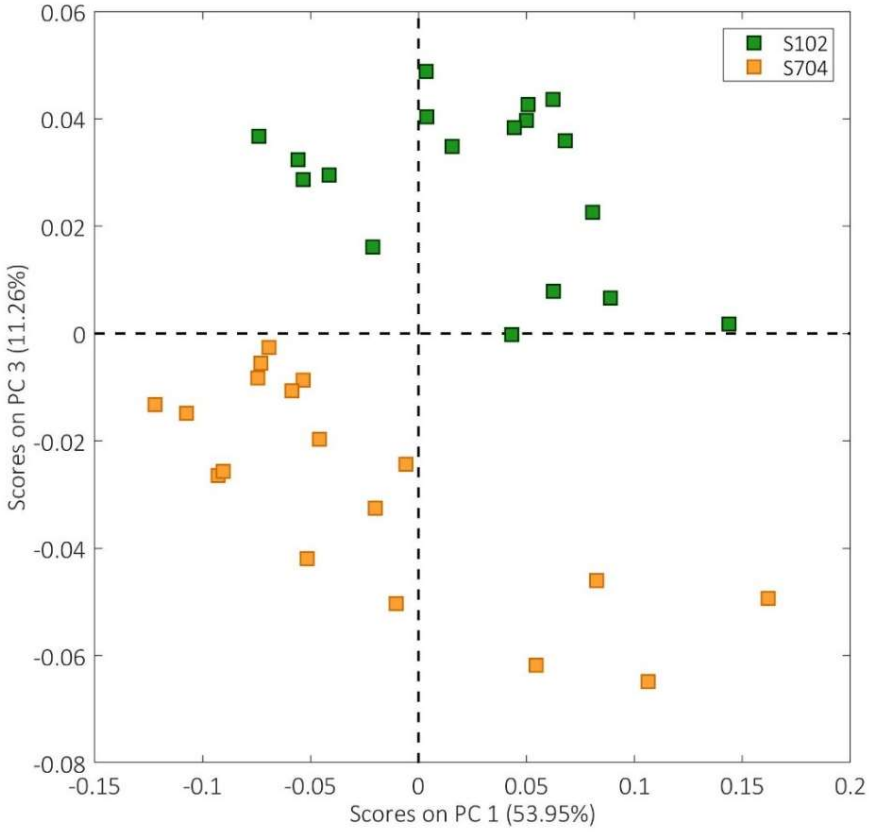

Figure 5-25. Scores plot of the collected spectral data of extra virgin olive oils with a different storage time (gap of six months of storage).

Taking into account only one tank (S704) and selecting the spectral region from 5400 cm^{-1} to 4900 cm^{-1}, related to H2O combinations, as the pre-processed spectra displayed differences in that region (Figure 5-26) a PCA model (Figure 5-27) has been developed accordingly.

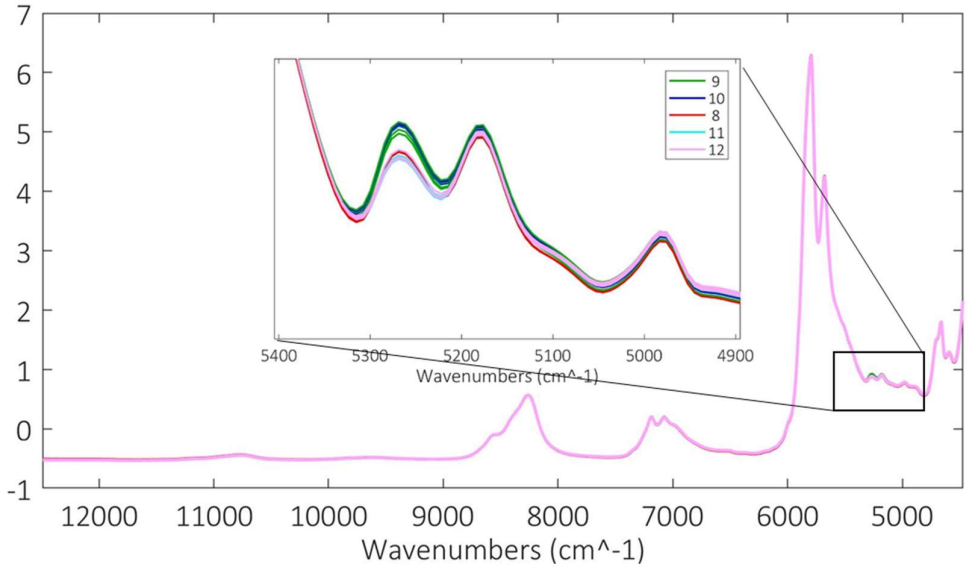

Figure 5-26. NIR spectral data resulting from extra virgin olive oil stored in tank S704 over the five months.

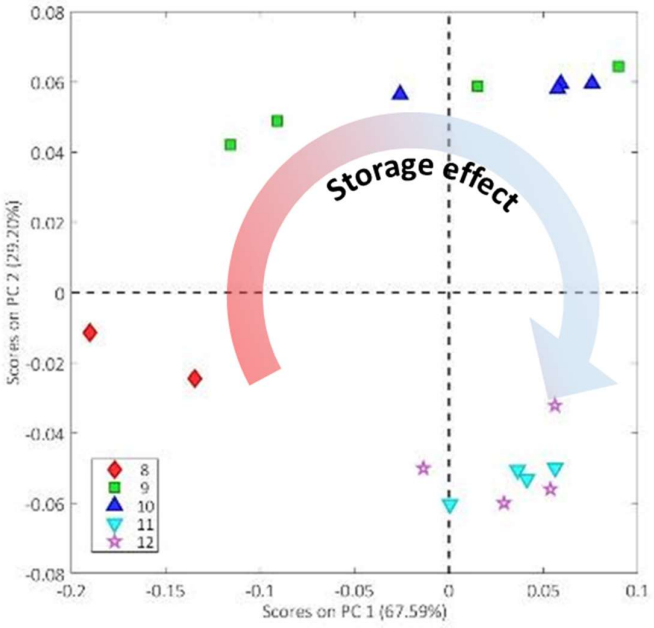

Figure 5-27. PCA scores plot of extra virgin olive oil during the five months of storage.

From the scores plot we can identify a distribution according to a storage effect related to the decreasing of the extra virgin olive oil temperature over the storage period of time as shown in Figure 5-28.

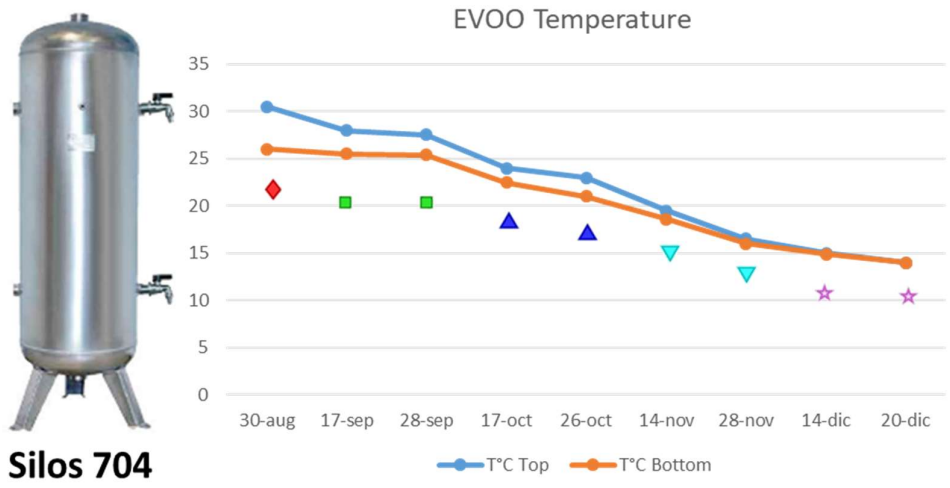

Figure 5-28. Evolution of the oil temperature, detected on the top and bottom, within tank S704.

Comparison between two cold pressing systems. Concerning the effect of pressing and experimental conditions, the spectral data arising from the FT-NIR testing of each oil have been subjected to exploratory data analysis providing PCA models for each oil, taking into account the two conditions. and an overall model including all the samples. From the PCA scores plot of sunflower, pumpkin and hemp vegetable oils (Figure 5-29) we can see that the samples are evenly distributed without distinction according to the two different pressing equipments and/or working conditions.

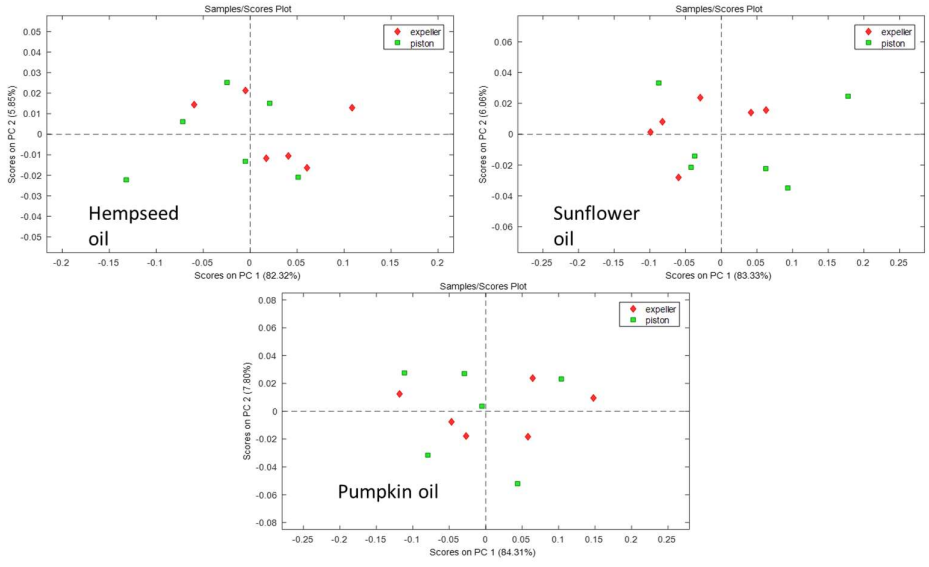

Figure 5-29. PCA models of hempseed, sunflower and pumpkin oils yielded starting from their seeds.

On the contrary, the effect of the two different presses is more obvious for the walnut oil and less evident for the linseed oil. Concerning the walnut and linseed oils the first principal component can split the samples depending on the two types of presses (Figure 5-30). Even for the abovementioned walnut and linseed samples there is less evidence about the effect of the two different conditions on the chemical composition of oils as the scores are close in the PCA model. The appearance of the walnut oil yielded from the two pressing systems was not the same, which reflected on the NIR spectra and accordingly on the PCA scores plot.

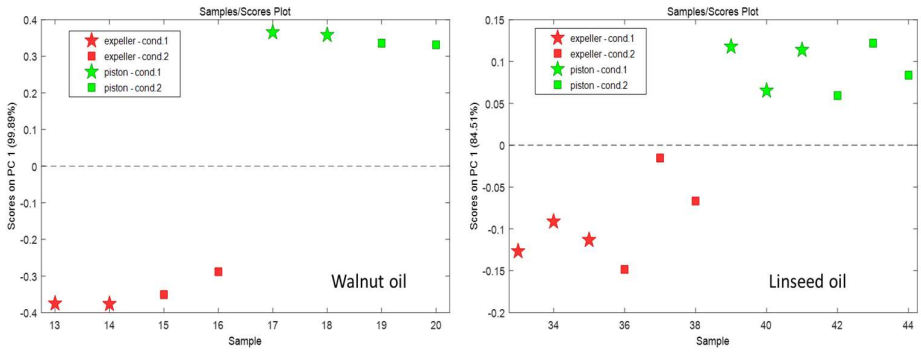

Figure 5-30. PCA models of walnut and linseed oils yielded by the two presses.

The overall PCA model displayed also a clear separation of walnut oil according to the two extraction tools whereas the distribution of linseed oil samples does not reflect the previous graphical depiction due to the augmented bi-dimensional space (Figure 5-31).

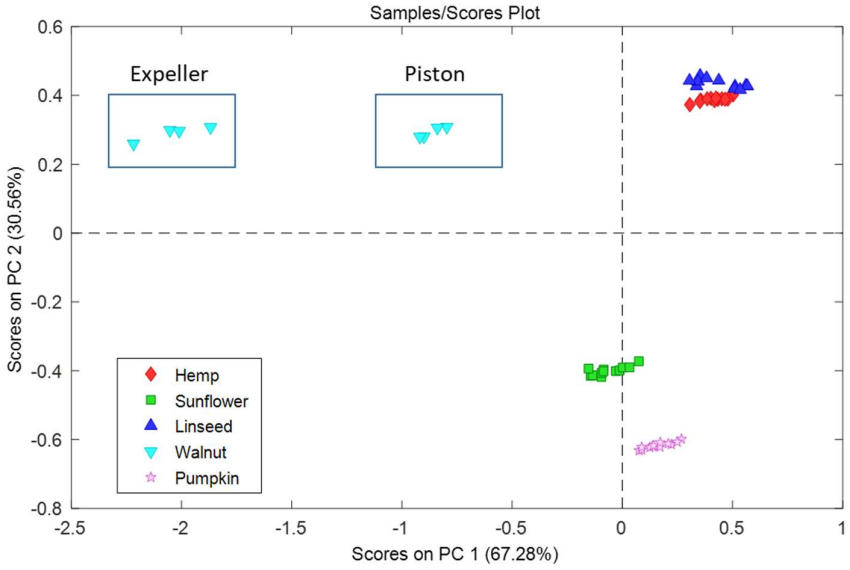

Figure 5-31. Overall PCA scores plot of the mean spectra of each replicate (three replicates for hemp, sunflower, linseed and pumpkin oils; two replicates for walnut oil).

Cold-pressed linseed oil oxidative stability. With respect to the employment of expeller press showed in Figure 5-19, the NIR spectra arising from the linseed oil before and after accelerated oxidation in stove taking also into account of the two extraction speeds show some differences in some specific regions (bands at 9000-8000 cm-1 due to C-H 2st overtone and bands at 7500-6000 cm-1) as displayed in Figure 5-32.

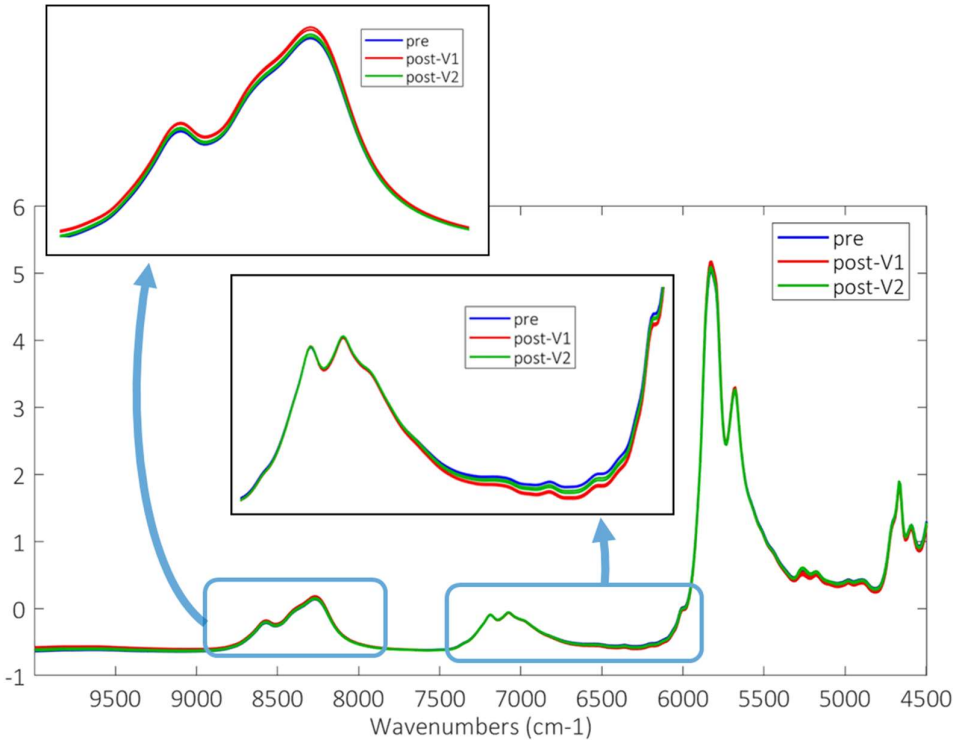

Figure 5-32. NIR spectra of linseed oil acquired by using the bench-top instrument, right after pressing under two different speeds, and after accelerated oxidation.

A PCA algorithm has been applied on the pre-processed spectral data. From the scores plot we can identify a limited area with close scores (blue samples) on the two principal components representing all the samples before oxidation. Following the oxidation treatment there is a wide distribution of the samples on scores plot according to one direction (red and green triangles). The oil samples (green triangles) arising from the higher pressing speed (cochlea speed = 80%) and subjected to accelerated oxidation are located in the upper part of the scores plot and distributed into two groups related to the two replicates (same pressing speed). On the other hand, the samples (red triangles) arising from the lower pressing speed (cochlea speed = 20%) are located towards the lower right part of the plot and divided into two groups due to the replicates (Figure 5-33).

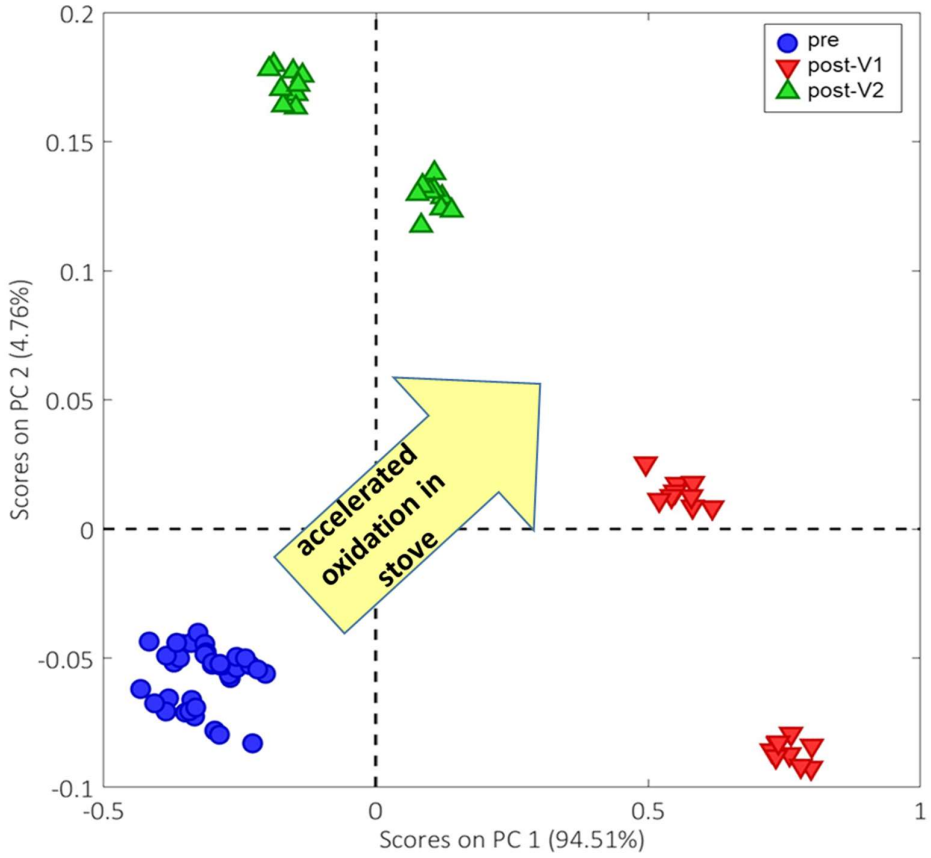

Figure 5-33. PCA scores plot of linseed oil samples before and after accelerated oxidation in stove.

5.2.4 References

Spatari, C., De Luca, M., Ioele, G., & Ragno, G. (2017). A critical evaluation of the analytical techniques in the photodegradation monitoring of edible oils. *LWT-Food Science and Technology*, *76*, 147-155.

Li, X., Kong, W., Shi, W., & Shen, Q. (2016). A combination of chemometrics methods and GC–MS for the classification of edible vegetable oils. *Chemometrics and Intelligent Laboratory Systems*, *155*, 145-150.

Dorni, C., Sharma, P., Saikia, G., & Longvah, T. (2018). Fatty acid profile of edible oils and fats consumed in India. *Food chemistry*, *238*, 9-15.

Armenta, S., Garrigues, S., & De la Guardia, M. (2007). Determination of edible oil parameters by near infrared spectrometry. *Analytica chimica acta*, *596*(2), 330-337.

Azizian, H., & Kramer, J. K. (2005). A rapid method for the quantification of fatty acids in fats and oils with emphasis on trans fatty acids using Fourier transform near infrared spectroscopy (FT-NIR). *Lipids*, *40*(8), 855-867.

Mignani, A. G., Ciaccheri, L., Thienpont, H., Ottevaere, H., Attilio, C., & Cimato, A. (2007, May). Toward a hyperspectral optical signature of extra virgin olive oil. In *Optical Sensing Technology and Applications* (Vol. 6585, p. 65852C). International Society for Optics and Photonics.

Armenta, S., Moros, J., Garrigues, S., & Guardia, M. D. L. (2010). The use of near-infrared spectrometry in the olive oil industry. *Critical reviews in food science and nutrition*, *50*(6), 567-582.

El-Hamidi, M., & Zaher, F. A. (2016). Comparison between some common clays as adsorbents of carotenoids, chlorophyll and phenolic compounds from vegetable oils. *American Journal of Food Technology*, *11*, 92-99.

Choe, E., & Min, D. B. (2006). Mechanisms and factors for edible oil oxidation. *Comprehensive reviews in food science and food safety*, *5*(4), 169-186.

Lorenzen, C. J., & Jeffrey, S. W. (1980). Determination of chlorophyll in seawater. *Unesco tech. pap. mar. sci*, *35*(1), 1-20.

5.3 Coffee industry

5.3.1 Introduction

The sensory features of coffee, one of the most popular and consumed beverage of the world, is significantly affected by the chemical composition of the green coffee beans, which is, in turn, highly correlated to their geographical origin. The variability existing among coffee grown in different world's regions has led, in many cases, to commercial frauds, like mislabelling and adulteration problems (Alonso-Salces, 2009). In this frame, the interest from coffee producers and industrial manufacturers in protecting the coffee market reputation has highly increased during the last decades and has led to the investigation of reliable techniques to assess the coffee authenticity. In this context, different analytical methods, such as liquid (LC) and gas chromatography (GC), mass spectrometry (MS) and Nuclear Magnetic Resonance (NMR) spectroscopy, were tested and allowed gathering accurate information about the coffee composition (Arana, 2015). However, these approaches present noticeable drawbacks, such as the time required, the cost, the complexity and the need of sample preparation before the analysis through the use of chemical solvents. A good solution to overcome the above mentioned issues is represented by NIR spectroscopy, which has widely demonstrated its suitability for the rapid and non-invasive prediction of important food quality parameters and has been successfully applied with different purposes in the coffee supply chain, such as the prediction of coffee species, the determination of organoleptic properties of the coffee beverage and the coffee colour after the roasting process (Bertone, 2016). Just a few studies (Marquetti,

2016; Medina, 2017) were instead found regarding the use of NIR spectroscopy for the direct determination of the geographical origin of coffee beans.

In this study, NIR spectroscopy, coupled with multivariate data analysis, was investigated for the development of a rapid and non-destructive method to classify green coffee beans based on their geographical origin. The FT-NIR spectra of coffee samples coming from different countries of Centre-South America and Asia were acquired and separately elaborated by two different laboratories. A preliminary data exploration was performed using the Principal Component Analysis (PCA). Subsequently, Partial Least Square-Discriminant Analysis (PLS-DA) classification models were developed by considering at first the continent and then the country of origin as discrimination parameter. Moreover, interval PLS-DA algorithm was investigated to select the most informative regions of NIR spectra.

The present study also considered the inter-laboratory comparison of the model results, which was performed using the McNemar test. A further inter-laboratory model validation was finally carried out by predicting the spectral test set of a laboratory using the model calibrated on the spectra collected by the other laboratory. The work described in this Section has been already published and can be found in Giraudo (2019).

The purpose of the study concerns the employment of NIRS coupled with chemometrics for the classification of green coffee beans based on their geographical origin.

5.3.2 Materials and methods

5.3.2.1 Coffee samples

An overall dataset of 191 different samples of green coffee beans was considered in this study. The first 88 samples came from countries belonging to the Centre-South America (Brazil, Honduras, Colombia, Costa Rica, Guatemala, Nicaragua), while the remaining 103 ones were harvested in Asian countries (India, Vietnam, Indonesia). The samples, each one consisting of three hundred grams of coffee beans, were vacuum-packed in light barrier packaging bags and send concurrently, within seven days from the preparation and in eight subsequent deliveries, to the two laboratories, i.e. the Department of Food, Environmental and Nutritional Sciences (UNIMI, University of Milan) and the Department of Applied Science and Technology (DISAT, Polytechnic of Turin).

5.3.2.2 Data collection

In both laboratories, the NIR spectra were collected using the same model of MPA FT-NIR spectrometer employed previously in Section 5.2, provided also with an integrating sphere for diffuse reflectance measurements. NIR spectra were acquired on the green coffee samples directly, i.e. without performing any kind of pre-treatment.

The acquisition parameters employed for spectra collection by FT-NIR are displayed as follows:

Acquisition modality: Absorbance
Spectral range: 12500 - 3600 cm^{-1}
Resolution: 8 cm^{-1}
Detector: InGaAs photodiodes
Sample scans: 64

Three replicate measurements were performed on each sample, keeping the sample holder in rotation during the spectra acquisition. All the measurements have been performed at room temperature and within 48 hours from the delivery of the coffee samples.

5.3.2.3 Multivariate data analysis

The data analysis has been independently performed on the NIR spectral data collected by the two laboratories using the PLS Toolbox running under MATLAB environment.

First, the replicate measurements of each samples have been averaged. Different pre-processing methods were then investigated, i.e. SNV, MSC, 1st and 2nd derivatives, combined in all cases with MC before the application of PCA algorithm.

For continent-based classification, PLS-DA models were built considering American and Asian samples as two independent classes. For the implementation of the country-based classification models five classes were instead considered: each one of the most representative countries in term of number of available samples (Brazil, Honduras, India, Vietnam) was labelled as an independent class, while all the remaining countries' samples, being their number not enough representative to build independent classes maintaining a balanced design, were grouped to form a unique class, named 'other'. The dataset was split into 75% training set and 25% validation set.

The performance of the PLS-DA classification models has been evaluated by comparing the reference class to the class predicted by the model. In this context, four different conditions may occur based on the model prediction, since the samples can result "true positive" (TP), "true negative" (TN), "false positive" (FP) or "false negative" (FN) (Szymanska, 2012; Ballabio, 2013). A series of statistical parameters were then calculated for all the classes separately i.e.:

- Sensitivity (SENS), which expresses the model capability to correctly recognize samples belonging to the considered class (Equation 5.1).
- Specificity (SPEC), which describes the model capability to correctly reject samples belonging to all the other classes (Equation 5.2).
- Efficiency (EFF), calculated as the geometric mean of SPEC and SENS (Equation 5.3).

$$\boldsymbol{SENS} = \frac{\boldsymbol{TP}}{(\boldsymbol{TP} + \boldsymbol{FN})} \qquad \text{Equation 5.1}$$

$$\boldsymbol{SPEC} = \frac{\boldsymbol{TN}}{(\boldsymbol{TN} + \boldsymbol{FP})} \qquad \text{Equation 5.2}$$

$$\boldsymbol{EFF} = \sqrt{(\boldsymbol{SPEC} * \boldsymbol{SENS})} \qquad \text{Equation 5.3}$$

All these parameters can assume values between 0 (0%) and 1 (100%) and were calculated referring to the calibration (TRN, CAL), to the cross-validation of the training set (TRN, CV), and to the prediction of the external test set (TST, PRED).

The variable selection was performed using the iPLS-DA algorithm, which consists, similarly to other interval-based chemometric techniques available in literature (Savorani, 2013), in cutting the whole spectral range in different intervals and building a series of 'local' models, first using one interval at a time and then adding the remaining intervals in an iterative way. The best model in terms of selected variables is the one characterized by the lowest Root Mean Square Error value achieved in cross-validation (RMSECV) (Leardi, 2004). In the present study, four different interval size values (i) were considered, with an arbitrary-defined length of 80, 40, 20 or 10 spectral variables, respectively.

Eventually, the achieved results were compared in terms of prediction performance applying the '*testcholdout*' Matlab function, which performs one-tailed, mid P-value McNemar test, a particular case of Fisher's sign test that verifies if two models have the same error rate (Grassi, 2018).

5.3.3 Results and discussion

PCA models. Both the PCA models calculated on the spectra collected from DISAT and POLITO laboratories allowed to appreciate a rather good separation of coffee samples after the combined application of SNV + MC as pre-treatment techniques. More than 80% of the total variance is explained by PC1, which is the main direction along which the coffee samples are separated according to both the continent and the country of origin (Figure 5-34a), while PC2 was found to essentially account for the within-class variability. As it can be clearly seen, only the samples belonging to the class 'other' did not grouped in a defined cluster. For this reason, they were excluded before the implementation of the country-based classification models.

A good correspondence was then found between the PCA loading values (data not shown) and the spectral wavelength regions responsible for the differences between American and Asian coffee samples (Figure 5-34b). In particular, negative loading values on PC1 were assigned to the wavelength regions where NIR signal is higher for the American samples (8450-8000 cm^{-1}, 7000 cm^{-1}, 5780 cm^{-1} and 5680 cm^{-1}), while positive loading values were assigned to the regions where NIR

signal is higher for the American samples (4020 cm^{-1} and 5000 cm^{-1}). The above mentioned spectral regions are essentially related to the NIR absorption of the C-H bonds of caffeine, amino- and fatty acids and lignin, the O-H bonds of cellulose and the N-H bonds of proteins and polyamides.

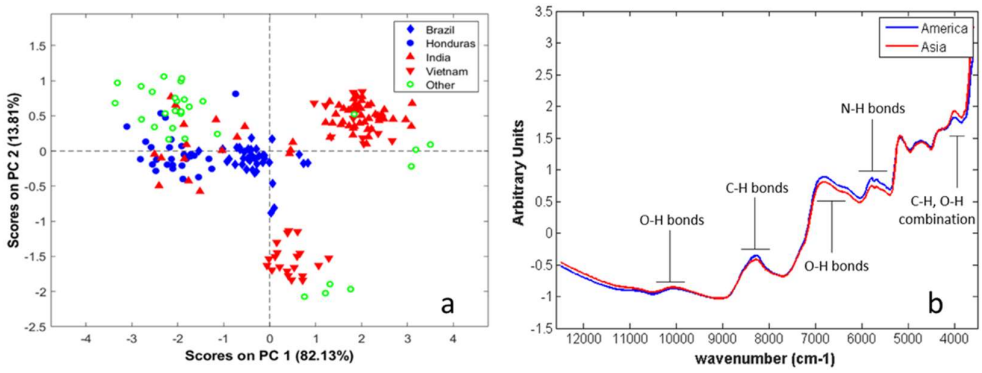

Figure 5-34. PCA scores plot of samples having different country of origin (a) and the average of SNV American and Asian NIR spectra of green coffee beans (b).

Classification models. The continent-based PLS-DA models developed using both the DISAT and POLITO spectra led to an EFF on the external test set (EFF_{TS}) always higher than 93%, no matter the considered data pre-treatment (Table 5-4).

Table 5-4. Results of continent-based Partial Least Squares-Discriminant Analysis (PLS-DA) on NIR spectra collected by DISAT and UNIMI after the application of different mathematical pre-processing: EFF obtained in calibration (CAL), cross-validation (CV) and prediction of the external test set (TS); Raw: raw spectral data without any pre-processing except MC; SNV: standard normal variate; MSC: multiplicative scatter correction; d1: first derivative; d2: second derivative.

		AMERICA			**ASIA**		
		CAL	***CV***	***TS***	***CAL***	***CV***	***TS***
DISAT	Raw	98.6	98.6	100.0	94.6	94.6	100.0
	SNV	98.6	98.6	100.0	93.2	93.2	93.1
	SNV + d1	98.6	98.6	100.0	93.2	93.2	96.6
	SNV + d2	98.6	98.6	100.0	95.9	93.2	96.6
	MSC	98.6	98.6	100.0	93.2	93.2	93.1
	MSC + d1	98.6	98.6	100.0	93.2	93.2	96.6
	MSC + d2	98.6	98.6	100.0	95.9	93.2	96.6
	d1	98.6	98.6	100.0	95.9	95.9	100.0
	d2	98.6	98.6	100.0	97.3	94.6	96.6
UNIMI	Raw	98.6	98.6	100.0	97.3	97.3	100.0
	SNV	98.6	98.6	100.0	93.2	91.9	96.5
	SNV + d1	98.6	98.6	100.0	94.6	93.2	93.1
	SNV + d2	97.2	97.2	100.0	94.6	93.2	96.5

MSC	98.6	98.6	100.0	93.2	91.9	96.5
MSC + d1	98.6	98.6	100.0	94.6	93.2	93.1
MSC + d2	97.2	97.2	100.0	94.6	93.2	96.5
d1	98.6	98.6	100.0	91.9	91.9	96.5
d2	97.2	97.2	100.0	95.9	93.2	96.5

Concerning the country-based PLS-DA models, up to the 98.1% of the Brazilian test set samples were correctly predicted, while from 90.0% to 97.8% of the Honduran, Indian and Vietnamese test set samples were properly assigned, depending on the pre-processing technique applied on the spectral dataset.

The iPLS-DA variable selection further improved the model results, with EFF_{TS} of 100% and 96.5% for the classification of American and Asian test set samples, respectively. Regarding the country-based classification, the iPLS-DA models led to EFF_{TS} ranging from 100% to 94.9%, depending on the class and the interval size of contiguous spectral variables considered.

In particular, the best iPLS-DA continent-based model was built using SNV + MC data pre-processing and just 40 spectral variables out of 1154 (which correspond to the whole spectral range) selected in four different spectral regions, i.e. 12258-12188 cm^{-1}, 5855-5786 cm^{-1}, 5315-5246 cm^{-1} and 4852-4783 cm^{-1}. The best performance in terms of country-based classification was always achieved on the SNV + MC pre-processed data, but selecting 90 spectral variables in three different spectral regions (9018-8871 cm^{-1}, 8632-8177 cm^{-1} and 6009-5940 cm^{-1}).

The McNemar test was always performed by comparing both the results achieved by DISAT and UNIMI separately and then their combination.

In all cases, at least the models calculated using the SNV and MSC pre-processed spectra were comparable ($P > 0.05$) in term of prediction performances. This similarity was further proven by the very good results achieved by the 'cross-laboratory' validation approach. As a matter of fact, the cross-laboratory models (i.e. DISAT calibration set - UNIMI external set and vice versa) led to correctly classify up to 100% of American and Asian test set samples and up to 95% of the test set samples according to the country of origin.

5.3.4 References

Alonso-Salces R. M., Serra F., Reniero F. and Héberger K. (2009). Botanical and Geographical Characterization of Green Coffee (Coffea arabica and Coffea canephora): Chemometric Evaluation of Phenolic and Methylxanthine Contents. *Journal of Agricultural and Food Chemistry*, 57(10), 4224-4235.

Arana V. A., Medina J., Alarcon R., Moreno E., Heintz L., Schäfer H. and Wist J. (2015). Coffee's country of origin determined by NMR: The Colombian case. *Food Chemistry*, 175, 500-506.

Bertone E., Venturello A., Giraudo A., Pellegrino G. and Geobaldo F. (2016). Simultaneous determination by NIR spectroscopy of the roasting degree and Arabica/Robusta ratio in roasted and ground coffee. *Food Control*, 59, 683-689.

Marquetti I., Link J. V., Guimarães Lemes A. L., dos Santos Scholz M. B., Valderrama P. and Bona, E. (2016). Partial least square with discriminant analysis and near infrared spectroscopy for evaluation of geographic and genotypic origin of Arabica coffee. *Computers and Electronics in Agriculture*, 121, 313-319.

Medina J., Caro Rodríguez D., Arana V. A., Bernal A., Esseiva P. and Wist J. (2017). Comparison of Attenuated Total Reflectance Mid-Infrared, Near Infrared, and 1H-Nuclear Magnetic Resonance Spectroscopies for the Determination of Coffee's Geographical Origin. *International Journal of Analytical Chemistry*, 1-8. 10.1155/2017/7210463.

Grassi S., Casiraghi E. and Alamprese C. (2018). Handheld NIR device: A non-targeted approach to assess authenticity of fish fillets and patties. *Food Chemistry*, 243, 382-388.

Chapter 6

Conclusions and future perspectives

6.1 Conclusions

6.1.1 Pharmaceutical industry

Raw materials. The compliance of raw materials with the quality standards is of crucial importance because of the intended use of the final products.

NIRS coupled with chemometrics showed the potential for rapid identity confirmation and/or classification of incoming materials, including botanical raw materials ensuring their quality based on the feed-forward control approach included in the PAT strategy.

The classification model developed and implemented without the employment of variable selection methods, related to ginkgo biloba dry extracts (botanical raw material), successfully predicted new samples of the three different classes which did not belong to the set of data used to build the model.

Concerning the composite raw material (soybean dry extract), the PCA model displayed a good separation of batches, with the same denomination, based on the two different suppliers.

Semi-finished and/or end product. The quantity of DHA in two final products was successfully predicted using the previously developed model.

The developed linear regression model, according to the evaluation criteria, i.e. RMSECV and R^2, and the number of principal components (4PCs), can be considered a good and simple model able to predict the active ingredient in end products. The abovementioned positive outcome showed the potential for the implementation of the quantitative prediction model within the dashboard enabling a fast outcome when it comes to predict the quantity of DHA in the semi-finished and end product, allowing in this way the company to make quick decisions about the production process.

This study showed the potential application of a quality control approach involving the several stages of processing, starting from the raw materials, semi-finished and end products as indicated by the PAT guidelines taking into account the complexity of quality control.

The reduction of the analytical time of response regarding incoming raw materials quality and semi-finished products conformity can improve the

management of resources (natural, human and financial) and manufacturing planning accordingly.

6.1.2 Vegetable oil company

Shelf-life assessment. The exploratory data analysis based on the visible spectral region of the three oils provided a clear separation of hemp oil subjected at the two different light sources (LED and NEON) after four months of exposure due to the different impact of light initiators (NEON showed a more drastic impact) on the chlorophylls, which are the main pigments of hemp oil effecting the sensory characteristics. The PCA model build on the NIR spectra did not provide a clear effect, based on the two various light sources, as the NIR spectral profile of hemp oil is almost the same over the light treatment period contrary to the visible spectra.

Extra virgin olive oil. The effect of the different storage period or cultivar that characterize the two olive oils is displayed with two separate groups on the PCA scores plot. The storage effect displayed on the scores plot of qualitative model arises from the decreasing trend of the oil temperature over the storage period of five months.

Comparison between two cold pressing systems. The two pressing tools have shown different effects depending on the type of seeds (walnut oil differentiates from the other oils based on the pressing system) while the impact of the two operating conditions was insignificant as showed by the PCA scores plot of the single vegetable oils and the overall model.

Cold-pressed linseed oil oxidative stability. The two different expeller press working conditions related to the pressing speed (20% and 80%) showed a similar outcome based on the PCA model stemming from the NIR spectra. On the contrary, the effect of accelerated oxidation on the linseed oil produced by setting two cochlea speeds is variable as showed by the qualitative model. The portable FT-NIR instrument equipped with a fibre optic transflectance probe showed the potential for in-line monitoring of the pressing process where the pressing speed affected the spectral signal to noise ratio.

6.1.3 Coffee industry

NIRS and chemometrics, through this study, proved their potential for the determination of the geographical origin of green coffee beans.

The similarity of results in calibration and validation (cross-validation and test set validation) revealed the robustness of the developed classification models referring either to the continent or country based approach.

The automation of the analysis based on the speed, objectivity and non-invasive nature is the main advantage of the proposed method.

The discrimination performances attained by the iPLS-DA algorithm have been superior than those obtained using the whole NIR spectral region. During the

development of handheld instruments in order to perform NIR analysis directly in field, the abovementioned advantage should be taken into account. Moreover, the proposed method can be used among several production sites or industries because of the successful cross-laboratory model validation.

6.2 Future perspectives

6.2.1 Pharmaceutical industry

Raw materials. The abovementioned platform, takin advantage of NIRS and multivariate data analysis, which results in a quick conformity response besides the advantage to display the bi-dimensional classification model can be extended to other raw materials of organization's interest once the number of batches is large enough to represent the variability of samples.

The quantitative prediction of active constituents in composite raw materials such as botanicals can also be performed in a rapid way, by linear regression methods (e. g. PLS algorithm) starting from the NIR spectra and the reference value of the interested component/s. This potential can enable a quick comparison with the ingredient's content stated in raw materials data sheet, instead of the employment of time consuming techniques.

Authenticity inspection and the classification of materials based on their origin can be achieved by the untargeted analytical technique (NIRS) coupled with chemometrics.

Semi-finished product. With respect to the compliance of semi-finished and final products to the specifications, quantitative models related to other products can be developed, based on the abovementioned approach, and subsequently implemented to achieve the organization's goals.

Regression models which can enable the simultaneous quantitative prediction of the various constituents of the product can be developed based for instance on the designs for multivariate calibration (illustrated in Chapter 3, Section 3.1.5).

6.2.2 Vegetable oil company

Shelf-life assessment. The assessment of the impact of LED and NEON on the three vegetable oils is in progress and will end after a year of light exposure.

Extra virgin olive oil. As the storage period of olive oil has been rather short, the extension of this interval of time can provide more information due to the temperature change over a longer period.

Comparison between two cold pressing systems. The impact of the two systems of pressing can be further assessed by using more pressing conditions according to an experimental design and further diversifying the seeds.

Cold-pressed linseed oil oxidative stability. The lapse of time and the conditions of accelerated oxidation can be varied in order to better understand the

process employing NIRS and chemometrics. An in-line testing system based on NIRS can be used for the monitoring of other pressing processes made with various presses in order to compare the qualitative outcome such as turbidity, particle suspensions, etc. Some qualitative parameters such as peroxide value, could be monitored during the process taking advantage of the prediction models previously developed.

6.2.3 Coffee industry

Model optimization and scale-up may be the next steps to perform taking into account a subsequent application as a tool for rapid conformity assessment of green coffee beans, within the feed-forward control strategy, before the roasting process.

The research study can be extended, for instance, by increasing the number of involved countries.